# Lesson 1

Big Data I.S.P.M.

<https://azure.microsoft.com/en-us/solutions/big-data/>

Big data: Massive amount of data which cannot be stored, processed and analyzed using the traditional ways.

Big data 5 V’s: **Volume, Veracity, Value, Velocity, Variety**

Advanced analytics on big data:

622

522

422

322

222

1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tool** | | **Ingest** | **Store** | **Prep & Train** | **Model & Serve** |
| Icon  Description automatically generated | **Azure Data Factory** | ✓ |  |  |  |
|  | **Azure Synapse Analytics**  **(Azure SQL DW)** | ✓ |  |  | ✓ |
| Logo, icon  Description automatically generated | **Azure Stream Analytics** | ✓ |  | ✓x` |  |
| Icon  Description automatically generated | **Azure Files** |  | ✓ |  |  |
| Icon  Description automatically generated | **Azure Queue** |  | ✓ |  |  |
| Icon  Description automatically generated | **Azure Data Lake Storage Gen 2**  (Non-relational data store) |  | ✓ |  |  |
| **A picture containing icon  Description automatically generated** | **CosmosDB**  (Non-relational data store) |  | ✓ |  | ✓ |
| Icon  Description automatically generated | **Azure Blob (WASB)**  (Non-relational data store) |  | ✓ |  |  |
| Logo  Description automatically generated | **Azure HDInsight**  **(Hadoop for azure)** |  |  | ✓ |  |
| **Icon  Description automatically generated** | **Azure Data Lake Analytics** |  |  | ✓ |  |
| **Icon  Description automatically generated** | **Azure Databricks**  **(python, Scala, Spark SQL, Spark R, Spark Structured Streaming)** |  |  | ✓ |  |
| **Icon  Description automatically generated** | **Power BI** |  |  |  | ✓ |

# Lesson 2

Data types & Storage

**Structured data**

- Structure is defined at design time.

- Data structure is designed in the form of tables.

**Semi-structured data.**

- Non-relational or NoSQL data (won’t fit neatly into tables, rows, and columns) – Uses tags or keys that organize and provide a hierarchy for the data

**Nonstructured data**

- Examples of nonstructured data include binary, audio, and image files

- The data structure is defined only when the data is read.

- Nonrelational systems can also support semistructured data such as JSON file formats.

- The schema of unstructured data is typically defined at query time. This means that data can be loaded onto a data platform in its native format.

Implement Relational Data Stores

The open-source world offers four types of NoSQL databases:

1. Key-value store: Stores key-value pairs of data in a table structure.
2. Document database: Stores documents that are tagged with metadata to aid document searches.
3. Graph database: Finds relationships between data points by using a structure that's composed of vertices and edges.
4. Column database: Stores data based on columns rather than rows. Columns can be defined at the query's runtime, allowing flexibility in the data that's returned performantly.

# Lesson 3

Data Storage

<https://docs.microsoft.com/en-us/learn/modules/intro-to-data-in-azure/3-how-azure-storage-meets-your-business-storage-needs>

[Azure] Four Storage options:

**Azure Blob (WASB):** A scalable object store for text and binary data

This is a data store that will store but not query data, your cheapest option is to set up a storage account as a Blob store.

Blob storage works well with images and unstructured data

Flexible pricing options (cold vs hot storage)

Icon

Description automatically generated

Icon

Description automatically generated

**Azure Files:** Managed file shares for cloud or on-premises deployments. Accessible via the industry standard Server Message Block (SMB) protocol

Icon

Description automatically generated

**Azure Queue:** Azure Queue storage is a service for storing large numbers of messages that can be accessed from anywhere in the world.

Azure Table: A NoSQL store for no-schema storage of structured data

Diagram, icon

Description automatically generated

Azure Storage Tables is aimed at high capacity on a single region (optional secondary read only region but no failover), indexing by PK/RK and storage optimized pricing;

Azure Cosmos DB is a globally distributed database service.

**high throughput** (single-digit millisecond latency),

**global distribution** (multiple failover), SLA-backed predictive performance with automatic indexing of each attribute/property and a pricing model focused on throughput.

**A picture containing icon

Description automatically generated**

Implement non-relational data stores:

- Azure Data Lake v2

- WASB (Blob storage)

- CosmosDB

[GE] A company plans to use Azure Storage for file storage purposes. Compliance rules require: - A single storage account to store all operations including reads, writes and deletes

- Retention of an on-premises copy of historical operations

- You need to configure the storage account.

Which two actions should you perform?

**A. Configure the storage account to log read, write and delete operations for service type Blob**

**B. Use the AzCopy tool to download log data from $logs/blob**

C. Configure the storage account to log read, write and delete operations for service-type table

D. Use the storage client to download log data from $logs/table

E. Configure the storage account to log read, write and delete operations for service type queue

Storage Logging logs request data in a set of blobs in a blob container named $logs in your storage account. You can also use the Azure Storage team provided command-line Azure Copy Tool (AzCopy) to download your log data.

[GE] You are developing a data engineering solution for a company. The solution will store a large set of key-value pair data by using Microsoft Azure Cosmos DB. The solution has the following requirements:

* Data must be partitioned into multiple containers.
* Data containers must be configured separately.
* Data must be accessible from applications hosted around the world.
* The solution must minimize latency.

Answer: Provision an Azure Cosmos DB account with the Azure Table API. Enable multi-region writes.

*NOT geo-redundancy*

[GE] You have an Azure subscription that contains an Azure Storage account. You plan to implement changes to a data storage solution to meet regulatory and compliance standards. Every day, Azure needs to identify and delete blobs that were NOT modified during the last 100 days. Solution: You apply an Azure Blob storage lifecycle policy. Does this meet the goal?

**Yes. Azure Blob storage lifecycle management offers a rich, rule-based policy for GPv2 and Blob storage accounts. Use the policy to transition your data to the appropriate access tiers or expire at the end of the data's lifecycle.**

<https://docs.microsoft.com/en-us/azure/storage/blobs/storage-lifecycle-management-concepts?tabs=azure-portal>

[GE] A company plans to use Azure SQL Database to support a mission-critical application. The application must be highly available without performance degradation during maintenance windows. You need to implement the solution. Which three technologies should you implement?

**A. Premium service tier**

B. Virtual machine Scale Sets

C. Basic service tier

D. SQL Data Sync

**E. Always On availability groups**

**F. Zone-redundant configuration**

Premium/business critical service tier model that is based on a cluster of database engine processes. This architectural model relies on a fact that there is always a quorum of available database engine nodes and has minimal performance impact on your workload even during maintenance activities.

E: In the premium model, Azure SQL database integrates compute and storage on the single node. High availability in this architectural model is achieved by replication of compute (SQL Server Database Engine process) and storage (locally attached SSD) deployed in 4-node cluster, using technology similar to SQL Server Always On Availability Groups.

F: Zone redundant configuration. By default, the quorum-set replicas for the local storage configurations are created in the same datacenter. With the introduction of Azure Availability Zones, you can place the different replicas in the quorum-sets to different availability zones in the same region. To eliminate a single point of failure, the control ring is also duplicated across multiple zones as three gateway rings (GW).

[GE] A company plans to use Azure Storage for file storage purposes. Compliance rules require:

- A single storage account to store all operations including reads, writes and deletes

- Retention of an on-premises copy of historical operations

- You need to configure the storage account.

Which two actions should you perform?

**A. Configure the storage account to log read, write and delete operations for service type Blob**

**B. Use the AzCopy tool to download log data from $logs/blob**

C. Configure the storage account to log read, write and delete operations for service-type table

D. Use the storage client to download log data from $logs/table

E. Configure the storage account to log read, write and delete operations for service type queue

[GE] You manage a financial computation data analysis process. Microsoft Azure virtual machines (VMs) run the process in daily jobs, and store the results in virtual hard drives (VHDs.) The VMs product results using data from the previous day and store the results in a snapshot of the VHD. When a new month begins, a process creates a new VHD.

You must implement the following data retention requirements:

- Daily results must be kept for 90 days

- Data for the current year must be available for weekly reports

- Data from the previous 10 years must be stored for auditing purposes

- Data required for an audit must be produced within 10 days of a request.

You need to enforce the data retention requirements while minimizing cost.

How should you configure the lifecycle policy? B {T,T,D} S



# Lesson 4

SQL

Azure SQL

Database elastic pools are a simple, cost-effective solution for managing and scaling multiple databases that have varying and unpredictable usage demands. The databases in an elastic pool are on a single server and share a set number of resources at a set price.

SQL on-premises

Data Sync is based around the concept of a sync group. A sync group is a group of databases that you want to synchronize.

|  |  |  |
| --- | --- | --- |
| Data sync tool | description | Disadvantages |
| transactional replication - | Transactional replication typically starts with a snapshot of the publication database objects and data. As soon as the initial snapshot is taken, subsequent data changes and schema modifications made at the Publisher are usually delivered to the Subscriber as they occur (in near real time). | Cannot publish from Azure SQL Database to on-prem |
| Data Migration Assistant (DMA) | synchronize data unidirectionally to from on-prem to Azure SQL Database |  |
| Azure SQL Data Sync | synchronize data between Azure SQL Database and any other SQL endpoints unidirectionally or bidirectionally. It enables hybrid SQL deployment and allows local data access from both Azure and on-premises application | High performance impact |
| backup and restore | Not a great method for sync |  |
| SQL Server Agent job | Not a great method for sync |  |

Storage can be configured between 5 GB and 4 TB with 1 GB increments. Storage is priced at GB/month.

[GE] A company is designing a hybrid solution to synchronize data and on-premises Microsoft SQL Server database to Azure SQL Database. You must perform an assessment of databases to determine whether data will move without compatibility issues. You need to perform the assessment. Which tool should you use?

A. SQL Server Migration Assistant (SSMA)

B. Microsoft Assessment and Planning Toolkit

C. SQL Vulnerability Assessment (VA)

D. Azure SQL Data Sync

**E. Data Migration Assistant (DMA)**

[GE] A company manages several on-premises Microsoft SQL Server databases. You need to migrate the databases to Microsoft Azure by using a backup process of Microsoft SQL Server. Which data technology should you use?

A. Azure SQL Database single database

B. Azure SQL Data Warehouse

C. Azure Cosmos DB

**D. Azure SQL Database Managed Instance**

[GE] You are a data architect. The data engineering team needs to configure a synchronization of data between an on-premises Microsoft SQL Server database to AzureSQL Database. Ad-hoc and reporting queries are being overutilized the on-premises production instance. The synchronization process must:

- Perform an initial data synchronization to Azure SQL Database with minimal downtime

- Perform bi-directional data synchronization after initial synchronization

You need to implement this synchronization solution. Which synchronization method should you use?

A. transactional replication

B. Data Migration Assistant (DMA)

C. backup and restore

D. SQL Server Agent job

**E. Azure SQL Data Sync**

[GE] A company uses Azure SQL Database to store sales transaction data. Field sales employees need an offline copy of the database that includes last year’s sales on their laptops when there is no internet connection available. You need to create the offline export copy. Which three options can you use?

**B. Export to a BACPAC file by using SQL Server Management Studio. Save the file to an Azure storage account**

**C. Export to a BACPAC file by using the Azure portal**

**E. Export to a BACPAC file by using the SqlPackage utility**

*The lesson here is do not attempt backups with any shell/cmd environments!*

[GE] A company has a SaaS solution that uses Azure SQL Database with elastic pools. The solution will have a dedicated database for each customer organization. Customer organizations have peak usage at different periods during the year. Which two factors affect your costs when sizing the Azure SQL Database elastic pools? Each correct answer presents a complete solution. NOTE: Each correct selection is worth one point.

**A. maximum data size**

B. number of databases

**C. eDTUs consumption**

D. number of read operations

E. number of transactions

[GE] A company has a SaaS solution that uses Azure SQL Database with elastic pools. The solution contains a dedicated database for each customer organization. Customer organizations have peak usage at different periods during the year. You need to implement the Azure SQL Database elastic pool to minimize cost. Which option or options should you configure?

A. Number of transactions only

B. eDTUs per database only

C. Number of databases only

D. CPU usage only

**E. eDTUs and max data size**

[GE] You have a container named Sales in an Azure Cosmos DB database. Sales has 120 GB of data. Each entry in Sales has the following structure. The partition key is set to the OrderId attribute. Users report that when they perform queries that retrieve data by ProductName, the queries take longer than expected to complete. You need to reduce the amount of time it takes to execute the problematic queries. Solution: You increase the Request Units (RUs) for the database. Does this meet the goal?

**Yes. To scale the provisioned throughput for your application, you can increase or decrease the number of RUs at any time**

[GE] You plan to use Microsoft Azure SQL Database instances with strict user access control. A user object must:

* Move with the database if it is run elsewhere
* Be able to create additional users

You need to create the user object with correct permissions. Which two Transact-SQL commands should you run?

A. ALTER LOGIN Mary WITH PASSWORD = 'strong\_password';

B. CREATE LOGIN Mary WITH PASSWORD = 'strong\_password';

**C. ALTER ROLE db\_owner ADD MEMBER Mary;**

**D. CREATE USER Mary WITH PASSWORD = 'strong\_password'**;

E. GRANT ALTER ANY USER TO Mary;

[GE] You have an Azure data solution that contains an Azure SQL data warehouse named DW1. Several users execute adhoc queries to DW1 concurrently. You regularly perform automated data loads to DW1. You need to ensure that the automated data loads have enough memory available to complete quickly and successfully when the adhoc queries run. What should you do?

A. Hash distribute the large fact tables in DW1 before performing the automated data loads.

**B. Assign a larger resource class to the automated data load queries.**

C. Create sampled statistics for every column in each table of DW1.

D. Assign a smaller resource class to the automated data load queries.

To ensure the loading user has enough memory to achieve maximum compression rates, use loading users that are a member of a medium or large resource class.

<https://docs.microsoft.com/en-us/azure/sql-data-warehouse/guidance-for-loading-data>

[GE] You have an Azure Storage account that contains 100 GB of files. The files contain text and numerical values. 75% of the rows contain description data that has an average length of 1.1 MB. You plan to copy the data from the storage account to an Azure SQL data warehouse. You need to prepare the files to ensure that the data copies quickly. Solution: You modify the files to ensure that each row is more than 1 MB. Does this meet the goal?

**No. Instead modify the files to ensure that each row is less than 1 MB**

[GE] You have an Azure Storage account that contains 100 GB of files. The files contain text and numerical values. 75% of the rows contain description data that has an

average length of 1.1 MB. You plan to copy the data from the storage account to an enterprise data warehouse in Azure Synapse Analytics. You need to prepare the files to ensure that the data copies quickly. Solution: You modify the files to ensure that each

row is less than 1 MB. Does this meet the goal?

No. Instead convert the files to compressed delimited text files. <https://docs.microsoft.com/en-us/azure/sql-data-warehouse/guidance-for-loading-data>

[GE] You have an Azure Storage account that contains 100 GB of files. The files contain text and numerical values. 75% of the rows contain description data that has an

average length of 1.1 MB. You plan to copy the data from the storage account to an Azure SQL data warehouse. You need to prepare the files to ensure that the data copies quickly. Solution: You copy the files to a table that has a columnstore index. Does this meet the goal?

**No. Instead modify the files to ensure that each row is less than 1 MB**

[GE] You plan to create a dimension table in Azure Data Warehouse that will be less than 1 GB. You need to create the table to meet the following requirements:

- Provide the fastest query time.

- Minimize data movement.

Which type of table should you use?

A. hash distributed

B. heap

C. replicated

**D. round-robin**

Usually common dimension tables or tables that doesn’t distribute evenly are good candidates for round-robin distributed table. Note: Dimension tables or other lookup tables in a schema can usually be stored as round-robin tables. Usually these tables connect to more than one fact tables and optimizing for one join may not be the best idea. Also usually dimension tables are smaller which can leave some distributions empty when hash distributed. Round-robin guarantees a uniform data distribution.

<https://blogs.msdn.microsoft.com/sqlcat/2015/08/11/choosing-hash-distributed-table-vs-round-robin-distributed-table-in-azure-sql-dw-service/>

[GE] You develop data engineering solutions for a company. The company has on-premises Microsoft SQL Server databases at multiple locations. The company must integrate data with Microsoft Power BI and Microsoft Azure Logic Apps. The solution must avoid single points of failure during connection and transfer to the cloud. The solution must also minimize latency. You need to secure the transfer of data between on-premises databases and Microsoft Azure. What should you do?

A. Install a standalone on-premises Azure data gateway at each location

B. Install an on-premises data gateway in personal mode at each location

C. Install an Azure on-premises data gateway at the primary location

**D. Install an Azure on-premises data gateway as a cluster at each location**

You can create high availability clusters of On-premises data gateway installations, to ensure your organization can access on-premises data resources used in Power BI reports and dashboards. Such clusters allow gateway administrators to group gateways to avoid single points of failure in accessing on-premises data resources. The Power BI service always uses the primary gateway in the cluster unless it’s not available. In that case, the service switches to the next gateway in the cluster, and so on.

<https://docs.microsoft.com/en-us/power-bi/service-gateway-high-availability-clusters>

[GE] You are developing the data platform for a global retail company. The company operates during normal working hours in each region. The analytical database is used once a week for building sales projections. Each region maintains its own private virtual network.

Building the sales projections is very resource intensive are generates upwards of 20 terabytes (TB) of data.

Microsoft Azure SQL Databases must be provisioned.

- Database provisioning must maximize performance and minimize cost

- The daily sales for each region must be stored in an Azure SQL Database instance

- Once a day, the data for all regions must be loaded in an analytical Azure SQL Database instance

You need to provision Azure SQL database instances.

How should you provision the database instances?

**1. Azure SQL Database elastic pools**

**2. Azure SQL Database Hyperscale**

A Hyperscale database supports up to 100 TB of data and provides high throughput and performance, as well as rapid scaling to adapt to the workload requirements.

[GE] You have an Azure SQL server named Server1 that hosts two development databases named DB1 and DB2. You have an administrative workstation that has an IP address of 192.168.8.8. The development team at your company has an IP addresses in the range of

192.168.8.1 to 192.168.8.5.

You need to set up firewall rules to meet the following requirements:

- Allows connection from your workstation to both databases.

- The development team must be able connect to DB1 but must be prevented from connecting to DB2.

- Web services running in Azure must be able to connect to DB1 but must be prevented from connecting to DB2.

Which three actions should you perform?

**A. Create a firewall rule on DB1 that has a start IP address of 192.168.8.1 and an end IP address of 192.168.8.5.**

B. Create a firewall rule on DB1 that has a start and end IP address of 0.0.0.0.

**C. Create a firewall rule on Server1 that has a start IP address of 192.168.8.1 and an end IP address of 192.168.8.5.**

D. Create a firewall rule on DB1 that has a start and end IP address of 192.168.8.8.

**E. Create a firewall rule on Server1 that has a start and end IP address of 192.168.8.8.**

[GE] You have a table named SalesFact in an Azure SQL data warehouse. SalesFact contains sales data from the past 36 months and has the following characteristics:  
- Is partitioned by month  
- Contains one billion rows  
- Has clustered columnstore indexes  
All the beginning of each month, you need to remove data SalesFact that is older than 36 months as quickly as possible.  
Which three actions should you perform in sequence in a stored procedure?

**[1] Create an empty table named SalesFact\_Work that has the same schema as SalesFact**

**[2] Switch the partition containing the stale data from SalesFact to SalesFact\_Work**

**[3] Drop the SalesFact\_Work Table**

Table partitions enable you to divide your data into smaller groups of data. In most cases, table partitions are created on a date column.

To switch partitions between two tables, you must ensure that the partitions align on their respective boundaries and that the table definitions match.

[GE] You develop data engineering solutions for a company.

You must integrate the company's on-premises Microsoft SQL Server data with Microsoft Azure SQL Database. Data must be transformed incrementally.

You need to implement the data integration solution.

Which tool should you use to configure a pipeline to copy data?

A. Use the Copy Data tool with Blob storage linked service as the source

B. Use Azure PowerShell with SQL Server linked service as a source

**C. Use Azure Data Factory UI with Blob storage linked service as a source**

D. Use the .NET Data Factory API with Blob storage linked service as the source

**Lesson 5**

Cosmos DB

<https://azure.microsoft.com/sv-se/blog/azure-cosmos-db-partitioning-design-patterns-part-1/>

Azure Cosmos DB indexes all data by default. If you try to query the data by by a clomun without a partition key (e.g. "LastName") , you will get the result, but it will cost you more request units (RU/s) because queries without partition key become fan-out queries.

Fan-out queries check all partitions, which will cost you extra RU/s and may affect the performance of your application. If you have a small number of partitions with less data, you may not perceive any significant side effects of fan-out queries, but when you start getting in high numbers of partitions and large amounts of data, fan-out queries can be detrimental to your applications.

One option is to have two more lookup collections PNR and “LastName” for the mapping of PNR to “UserId”, and “LastName” to “UserId”. The PNR collection will have PNR as the partition key and row key and “UserId” as the value

**Cassandra API**

Cassandra takes a column-based approach to querying data. It is an open-source distributed NoSQL database.  Cassandra delivers outstanding performance due to its [distributed node design](https://www.datastax.com/dev/blog/2012-in-review-performance)

The Cassandra API enables you to interact with data stored in Azure Cosmos DB using the Cassandra Query Language (CQL) , Cassandra-based tools (like cqlsh) and Cassandra client drivers that you're already familiar with.

[GE] You have a container named Sales in an Azure Cosmos DB database. Sales has 120 GB of data. Each entry in Sales has the following structure. The partition key is set to the OrderId attribute. Users report that when they perform queries that retrieve data by ProductName, the queries take longer than expected to complete. You need to reduce the amount of time it takes to execute the problematic queries.

Solution: You create a lookup collection that uses ProductName as a partition key. Does this meet the goal?

*This needs a partition key AND a Value!!*

**No. One option is to have a lookup collection “ProductName” for the mapping of “ProductName” to “OrderId”**

[GE] You have a container named Sales in an Azure Cosmos DB database. Sales has 120 GB of data. Each entry in Sales has the following structure. The partition key is set to the OrderId attribute. Users report that when they perform queries that retrieve data by ProductName, the queries take longer than expected to complete. You need to reduce the amount of time it takes to execute the problematic queries.

Solution: You create a lookup collection that uses ProductName as a partition key and OrderId as a value. Does this meet the goal?

**Yes. One option is to have a lookup collection “ProductName” for the mapping of “ProductName” to “OrderId”**

[GE] You have an Azure Cosmos DB database that uses the SQL API. You need to delete stale data from the database automatically. What should you use?

A. soft delete

B. Low Latency Analytical Processing (LLAP)

C. schema on read

**D. Time to Live (TTL)**

With Time to Live or TTL, Azure Cosmos DB provides the ability to delete items automatically from a container after a certain period. By default, you can set time to live at the container level and override the value on a per-item basis. After you set the TTL at a container or at an item level, Azure Cosmos DB will automatically remove these items after the period, since the time they were last modified.

<https://docs.microsoft.com/en-us/azure/cosmos-db/time-to-live>

Distributed databases that rely on replication for high availability, low latency, or both, must make a fundamental tradeoff between the read consistency, availability, latency, and throughput as defined by the [PACLC theorem](https://en.wikipedia.org/wiki/PACELC_theorem).

Table

Description automatically generated

The consistency levels [S.B.S.C.E] are region-agnostic and are guaranteed for all operations regardless of the region from which the reads and writes are served.

**Strong**: Strong consistency offers a linearizability guarantee. Linearizability refers to serving requests concurrently. The reads are guaranteed to return the most recent committed version of an item. A client never sees an uncommitted or partial write. Users are always guaranteed to read the latest committed write.

**Bounded staleness**: The reads are guaranteed to honor the consistent-prefix guarantee. The reads might lag writes by at most "K" versions (that is, "updates") of an item or by "T" time interval, whichever is reached first.

**Session**: Within a single client session reads are guaranteed to honor the consistent-prefix, monotonic reads, monotonic writes, read-your-writes, and write-follows-reads guarantees. This assumes a single "writer" session or sharing the session token for multiple writers.

**Consistent prefix**: Updates that are returned contain some prefix of all the updates, with no gaps. Consistent prefix consistency level guarantees that reads never see out-of-order writes.

**Eventual**: There's no ordering guarantee for reads. In the absence of any further writes, the replicas eventually converge.  
Eventual consistency is the weakest form of consistency because a client may read the values that are older than the ones it had read before. Eventual consistency is ideal where the application does not require any ordering guarantees. Examples include count of Retweets, Likes, or non-threaded comments

[GE] You plan to implement an Azure Cosmos DB database that will write 100,000 JSON every 24 hours. The database will be replicated to three regions. Only one region will be writable. You need to select a consistency level for the database to meet the following requirements:

- Guarantee monotonic reads and writes within a session.

- Provide the fastest throughput.

- Provide the lowest latency. Which consistency level should you select?

A. Strong

B. Bounded Staleness

C. Eventual

**D. Session**

E. Consistent Prefix

Session: Within a single client session reads are guaranteed to honor the consistent-prefix (assuming a single “writer” session), monotonic reads, monotonic writes, read-your-writes, and write-follows-reads guarantees. Clients outside of the session performing writes will see eventual consistency

<https://docs.microsoft.com/en-us/azure/cosmos-db/consistency-levels-choosing>

[GE] You plan to deploy an Azure Cosmos DB database that supports multi-master replication. You need to select a consistency level for the database to meet the following requirements:

- Provide a recovery point objective (RPO) of less than 15 minutes.

- Provide a recovery time objective (RTO) of zero minutes.

What are three possible consistency levels that you can select?

A. Strong

B. Bounded Staleness

**C. Eventual**

**D. Session**

**E. Consistent Prefix**

*We don’t require great consistency. We only require low latency and high availability!*

**Lesson 5**

Units

**DTU** stands for Database Transaction Unit. DTUs give you a way to compare database performance across the service tiers offered by Azure. DTUs roughly measure performance as a combination of CPU, Memory, Reads, and Writes. When provisioning compute for elastic pools, the acronym eDTU may be used to refer to DTUs that are part of an elastic pool.

**RU** stands for Request Unit. The cost of all database operations is normalized by Azure Cosmos DB and is expressed by Request Units (or RUs, for short). You can think of RUs per second as the currency for throughput. RUs per second is a rate-based currency. It abstracts the system resources such as CPU, IOPS, and memory that are required to perform the database operations supported by Azure Cosmos DB

**SU** Stands for Streaming Units. These the computing resources that are allocated to execute a Stream Analytics job. The higher the number of SUs, the more CPU and memory resources are allocated for your job.

<https://docs.microsoft.com/bs-cyrl-ba/azure/stream-analytics/stream-analytics-monitoring>

**DIU** stands for Data Integration Unit (DIU) is a measure that represents the power of a single unit in Azure Data Factory. Power is a combination of CPU, memory, and network resource allocation. DIU only applies to [Azure integration runtime](https://docs.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime#azure-integration-runtime). DIU does not apply to [self-hosted integration runtime](https://docs.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime#self-hosted-integration-runtime).

**DWU** stands for Data Warehouse Unit. A [dedicated SQL pool (formerly SQL DW)](https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-overview-what-is) represents a collection of analytic resources that are being provisioned. Analytic resources are defined as a combination of CPU, memory, and IO. These three resources are bundled into units of compute scale called Data Warehouse Units (DWUs). A DWU represents an abstract, normalized measure of compute resources and performance.

[GE] Which two metrics should you use to identify the appropriate RU/s for the telemetry data?

**A. Number of requests**

B. Number of requests exceeded capacity

C. End to end observed read latency at the 99th percentile

D. Session consistency

**E. Data + Index storage consumed**

F. Avg Throughput/s

Scenario: The telemetry data must be monitored for performance issues. You must adjust the Cosmos DB Request Units per second (RU/s) to maintain a performance SLA while minimizing the cost of the RU/s. With Azure Cosmos DB, you pay for the throughput you provision and the storage you consume on an hourly basis. While you estimate the number of RUs per second to provision, consider the following factors:

Item size: As the size of an item increases, the number of RUs consumed to read or write the item also increases. Monitor and optimize data solutions

# Lesson 6

Storage Redundancy

[Azure] Storage Redundancy

|  |  |  |
| --- | --- | --- |
| **Abbr** | **Name** | **Description** |
| LRS | Locally Redundant Storage | Replicates your data three times within a single data center |
| ZRS | Zone-Redundant Storage | Replicates your data across three storage clusters in a single region. |
| GRS | Geo-Redundant Storage | Replicates your data to a secondary region. Can withstand regional outage. |
| RA-GRS | Read-Access Geo Redundant Storage | Provides read-only access to the data in the secondary location, in addition to GRS. |
| GZRS | Geo-Zone-Redundant Storage | Replicates data across three Azure Availability Zones in two regions. |
| RA-GZRS | Read-Access Geo Zone Redundant Storage | Provides read-only access to the data in the secondary location, in addition to GZRS. |

[Azure] Storage Redundancy

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Abbr** | LRS | ZRS | GRS | GZRS |
| Node Unavailability (within a DC) | Yes | Yes | Yes | Yes |
| DC outage |  | Yes | Yes | Yes |
| Region Outage |  |  | Yes | Yes |
| Read-Access in event of a Region Outage |  |  | RA-GRS | RA-GZRS |
| Durability over a given year | 11 9’s | 12 9’s | 16 9’s | 16 9’s |

# Lesson 7

Data Platforms

[Azure] Azure Cosmos DB is a globally distributed, multi-model database. You can deploy it by using several API models:

SQL API MongoDB API Cassandra API Gremlin API Table API

Because of the multi-model architecture of Azure Cosmos DB, you benefit from each model's inherent capabilities. For example, you can use MongoDB for semi-structured data, Cassandra for wide columns, or Gremlin for graph databases. Because of the multi-model architecture of Azure Cosmos DB, you benefit from each model's inherent capabilities. For example, you can use MongoDB for semi-structured data, Cassandra for wide columns, or Gremlin for graph databases.

[Azure] Azure SQL Database

**Use T-SQL to query the contents of a SQL Database.**

*Remember this is Azure SQL Database PaaS not an instance installed within Azure virtual machine*

Use SQL Database when you need to scale up and scale down OLTP systems on demand. SQL Database is a good solution when your organization wants to take advantage of Azure security and availability features

Data Services

<https://docs.microsoft.com/en-us/learn/modules/survey-the-azure-data-platform/10-azure-other-data-platform-services>

Logo, company name

Description automatically generated

[Azure] Databricks

Serverless platform that's optimized for Azure. One-click setup

Streamlined workflows

Interactive workspace + fully managed Spark clusters for Spark-based applications.

In Databricks notebooks you'll use familiar programming tools such as R, Python, Scala, and SQL

Icon

Description automatically generated

[Azure] Data Factory

**Organize raw data into meaningful data stores and data lakes**

Orchestrates the movement of data between various data stores.

Cloud-integration service.

Streamlined workflows

processes and transforms data by using compute services such as Azure HDInsight, Hadoop, Spark, and Azure Machine Learning

Graphical user interface, application, Teams

Description automatically generated

[Azure] Data Catalog

**Discover, understand, and consume data sources**

Is a fully managed cloud service

Is the best choice to store documentation about a data source

# Lesson 8

The Data Engineering Process

<https://docs.microsoft.com/en-us/learn/modules/data-engineering-processes/3-data-engineering-practices>

**ETL**

[Azure] Azure Data Factory v2



**ELT**

[Azure] Azure Data Factory v2

[Azure] Azure Synapse

[Azure] HDInsight with Hive

[Azure] Oozie on HDInsight

SQL Server Integration Services (SSIS)



**Extract**

During the extraction process, data engineers define the data and its source:

[1] Define the data source: Identify source details such as the resource group, subscription, and identity information such as a key or secret.

[2] Define the data: Identify the data to be extracted. Define data by using a database query, a set of files, or an Azure Blob storage name for blob storage.

**Transform**

[3] Define the data transformation: Data transformation operations can include splitting, combining, deriving, adding, removing, or pivoting columns. Map fields between the data source and the data destination. You might also need to aggregate or merge data.

**Load**

[4] Define the destination: During a load, many Azure destinations can accept data formatted as a JavaScript Object Notation (JSON), file, or blob. You might need to write code to interact with application APIs.

Azure Data Factory offers built-in support for Azure Functions. You'll also find support for many programming languages, including Node.js, .NET, Python, and Java. Although Extensible Markup Language (XML) was common in the past, most systems have migrated to JSON because of its flexibility as a semistructured data type.

[5] Start the job: Test the ETL job in a development or test environment. Then migrate the job to a production environment to load the production system.

[6] Monitor the job: ETL operations can involve many complex processes. Set up a proactive and reactive monitoring system to provide information when things go wrong. Set up logging according to the technology that will use it.

# Lesson 9

INGESTION: Data Ingestion

<https://docs.microsoft.com/en-us/learn/modules/explore-data-ingestion-azure/2-describe-common-practices-for-data-loading>

Data ingestion is the first part of any data warehousing solution. It is arguably the most important part. In a big data system, data ingestion has to be fast enough to capture the large quantities data that may be heading your way, and have enough compute power to process this data in a timely manner.

Ways to ingest data:

**Azure Data Factory v2** [Azure] **ELT + ETL**



**PolyBase** [Azure]

*Azure SQL Database does not support PolyBase*

**SQL Server Integration Services ELT + ETL**

Logo

Description automatically generated

**Icon

Description automatically generated**

**Azure Databricks** [Azure]



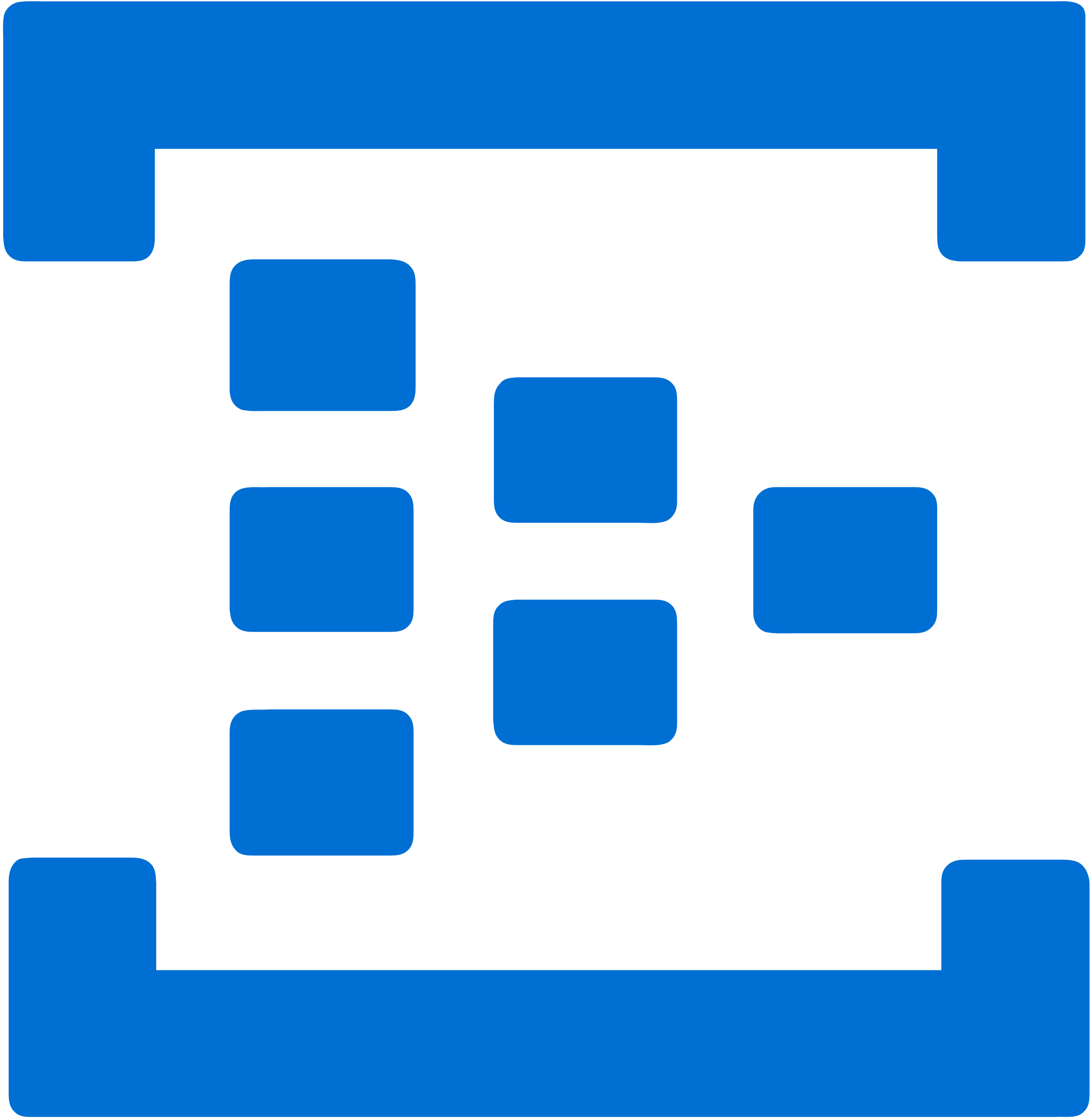
**Azure Synapse Analytics** [Azure]

Icon

Description automatically generated

**IoT Hub** [Azure]

Communication: device -> cloud & cloud -> device



**Event Hub** [Azure]

# Lesson 10

INGESTION: Azure Data Factory

Icon

Description automatically generated**Azure Data Factory v2** [Azure] **ELT + ETL**

Data Factory contains a series of interconnected systems that provide a complete end-to-end platform for data engineers.

As it ingests the data, Data Factory can clean, transform, and restructure the data, before loading it into a repository such as a data warehouse. Once the data is in the data warehouse, you can analyze it.

Data Factory provides an orchestration engine. Orchestration is the process of directing and controlling other services, and connecting them together, to allow data to flow between them

Diagram

Description automatically generated

**PolyBase** [Azure]

*Azure SQL Database does not support PolyBase*

Feature of SQL Server and Azure Synapse Analytics

Enables you to run Transact-SQL queries that read data from external data sources (makes these sources appear like SQL tables).

Data Factory can directly invoke PolyBase on your behalf if your data is in a PolyBase-compatible data store.

Logo

Description automatically generated**SQL Server Integration Services ELT + ETL**

*SSIS is an on-premises utility.*

use SSIS to solve complex business problems by copying or downloading files, loading data warehouses, cleaning and mining data, and managing SQL database objects and data. SSIS is part of Microsoft SQL Server.

How to ingest data from Azure SQL database into ADLS gen 2 using Data Factory:

2

Pre-requisite: ADLS gen 2

(should have the following configured)

Storage account + + +

Images adlsfactory adlsserver

A picture containing icon

Description automatically generatedIcon

Description automatically generatedIcon

Description automatically generated

NB: remember to toggle firewall rules in [Azure] SQL server

“Allow Azure services and resources to access this server”

Graphical user interface, text, application, chat or text message

Description automatically generated

A picture containing icon

Description automatically generated

[1] Add new linked services:

Manage > Linked Services > New> Azure SQL Database

-- test connection

Manage > Linked Services > New> Azure Data Lake Storage Gen2

-- test connection

Icon

Description automatically generated

Images

Author and Monitor Data Factory [Azure]

*NB: This can have issues. Try using incognito mode in your browser*

[2] Add datasets

*You should have a table already created in your db*

-- Add new dataset

Author > Azure SQL Database

-- Add new dataset

Author > Azure Data Lake Storage Gen2

Icon

Description automatically generated

[3] Create data Pipeline

-- Create “copy data” activity

-- Map between source and destination

-- Trigger pipeline

**Integration runtime**

APPLIES TO: Azure Data Factory | Azure Synapse Analytics

The Integration Runtime (IR) is the compute infrastructure used by Azure Data Factory to provide the following data integration capabilities across different network environments:

**How do I know what runtime to use?**

If your data store is located inside an on-premises network, an Azure virtual network, or Amazon Virtual Private Cloud, you need to configure a self-hosted integration runtime to connect to it.

Alternatively, if your data store is a managed cloud data service, you can use Azure integration runtime.

|  |  |  |
| --- | --- | --- |
| **IR type** | **Example** | **Technology used** |
| Azure | Storage Account (blob) -> Azure DW | Data Factory |
| Self-hosted | SQL -> Azure Data Lake Storage | Data Factory (V2 copy activity) |

Data Factory offers three types of Integration Runtime (IR), and you should choose the type that best serve the data integration capabilities and network environment needs you're looking for. These three types are:

|  |  |  |
| --- | --- | --- |
| **IR type** | **Public network** | **Private network** |
| Azure | Data Flow Data movement Activity dispatch | Data Flow Data movement Activity dispatch |
| Self-hosted | Data movement Activity dispatch | Data movement Activity dispatch |
| Azure-SSIS | SSIS package execution | SSIS package execution |

[GE] You have an Azure Storage account and an Azure SQL data warehouse in the UK South region. You need to copy blob data from the storage account to the data warehouse by using Azure Data Factory. The solution must meet the following requirements:

- Ensure that the data always remains in the UK South region.

- Minimize administrative effort. Which type of integration runtime should you use?

**A. Azure integration runtime**

B. Self-hosted integration runtime

C. Azure-SSIS integration runtime

*The DW is not on premises. Self-hosted integration runtime is to be used On-premises.*

[GE] You have an Azure virtual machine that has Microsoft SQL Server installed. The server contains a table named Table1. You need to copy the data from Table1 to an Azure Data Lake Storage Gen2 account by using an Azure Data Factory V2 copy activity. Which type of integration runtime should you use?

A. Azure integration runtime

**B. Self-hosted integration runtime**

C. Azure-SSIS integration runtime

Copying between a cloud data source and a data source in private network: if either source or sink linked service points to a self-hosted IR, the copy activity is executed on that self-hosted Integration Runtime.

<https://docs.microsoft.com/en-us/azure/data-factory/concepts-integration-runtime#determining-which-ir-to-use>

[GE] Your company has on-premises Microsoft SQL Server instance. The data engineering team plans to implement a process that copies data from the SQL Server instance to Azure Blob storage. The process must orchestrate and manage the data lifecycle. You need to configure Azure Data Factory to connect to the SQL Server instance.

Which three actions should you perform in sequence?

**[1] Deploy an Azure Data Factory**

**[2] From the on-premises network, install and configure a self-hosted runtime**

**[3] Configure a linked service to connect to the SQL Server instance**

[GE] You develop data engineering solutions for a company. You must integrate the company’s on-premises Microsoft SQL Server data with Microsoft Azure SQL Database. Data must be transformed incrementally. You need to implement the data integration solution. Which tool should you use to configure a pipeline to copy data?

A. Use the Copy Data tool with Blob storage linked service as the source

B. Use Azure PowerShell with SQL Server linked service as a source

**C. Use Azure Data Factory UI with Blob storage linked service as a source**

D. Use the .NET Data Factory API with Blob storage linked service as the source

The Integration Runtime is a customer managed data integration infrastructure used by Azure Data Factory to provide data integration capabilities across different network environments. A linked service defines the information needed for Azure Data Factory to connect to a data resource. We have three resources in this scenario for which linked services are needed:

On-premises SQL Server, Azure Blob Storage, Azure SQL database. A pipeline is a logical grouping of Activities, each of which defines the actions to perform on the data contained in Datasets. Linked services are used to define the information needed for Data Factory to connect to the data resources

[GE] You are monitoring the Data Factory pipeline that runs from Cosmos DB to SQL Database for Race Central. You discover that the job takes 45 minutes to run. What should you do to improve the performance of the job?

A. Decrease parallelism for the copy activities.

**B. Increase that data integration units.**

C. Configure the copy activities to use staged copy.

D. Configure the copy activities to perform compression.

Performance tuning tips and optimization features. In some cases, when you run a copy activity in Azure Data Factory, you see a "Performance tuning tips" message on top of the copy activity monitoring, as shown in the following example. The message tells you the bottleneck that was identified for the given copy run. It also guides you on what to change to boost copy throughput. The performance tuning tips currently provide suggestions like:

- Use PolyBase when you copy data into Azure SQL Data Warehouse.

- Increase Azure Cosmos DB Request Units or Azure SQL Database DTUs (Database Throughput Units) when the resource on the data store side is the

bottleneck.

Remove the unnecessary staged copy.

[GE] You need to ensure that phone-based poling data can be analyzed in the PollingData database. How should you configure Azure Data Factory?

A. Use a tumbling schedule trigger

B. Use an event-based trigger

C. Use a schedule trigger

D. Use manual execution

When creating a schedule trigger, you specify a schedule (start date, recurrence, end date etc.) for the trigger, and associate with a Data Factory pipeline. Scenario:

- All data migration processes must use Azure Data Factory.

- All data migrations must run automatically during non-business hours

[GE] Each day, company plans to store hundreds of files in Azure Blob Storage and Azure Data Lake Storage. The company uses the parquet format. You must develop a pipeline that meets the following requirements:

- Process data every six hours

- Offer interactive data analysis capabilities

- Offer the ability to process data using solid-state drive (SSD) caching

- Use Directed Acyclic Graph (DAG) processing mechanisms

- Provide support for REST API calls to monitor processes

- Provide native support for Python

- Integrate with Microsoft Power BI

You need to select the appropriate data technology to implement the pipeline. Which data technology should you implement?

A. Azure SQL Data Warehouse

**B. HDInsight Apache Storm cluster**

C. Azure Stream Analytics

D. HDInsight Apache Hadoop cluster using MapReduce

E. HDInsight Spark cluster

Storm runs topologies instead of the Apache Hadoop MapReduce jobs that you might be familiar with. Storm topologies are composed of multiple components that are arranged in a directed acyclic graph (DAG). Data flows between the components in the graph. Each component consumes one or more data streams and can optionally emit one or more streams. Python can be used to develop Storm components.

Spark: data parallel processing, more efficient than Hadoop MapReduce

Storm: task parallel computing, with independent workflows topologies (DAG’s) which execute until disturbance or the system shuts down completely.

[GE] You develop a data ingestion process that will import data to a Microsoft Azure SQL Data Warehouse. The data to be ingested resides in parquet files stored in an Azure Data Lake Gen 2 storage account. You need to load the data from the Azure Data Lake Gen 2 storage account into the Azure SQL Data Warehouse.

Solution:

1. Use Azure Data Factory to convert the parquet files to CSV files

2. Create an external data source pointing to the Azure storage account

3. Create an external file format and using the external data source

external table

4. Load the data using the INSERT–SELECT statement Does the solution meet the goal?

**No. There is no need to convert the parquet files to CSV files. You load the data using the CREATE TABLE AS SELECT statement**

[GE] You develop a data ingestion process that will import data to a Microsoft Azure SQL Data Warehouse. The data to be ingested resides in parquet files stored in an Azure Data Lake Gen 2 storage account. You need to load the data from the Azure Data Lake Gen 2 storage account into the Azure SQL Data Warehouse.

Solution:

1. Create an external data source pointing to the Azure storage account

2. Create an external file format and external table using the external data source

3. Load the data using the INSERT–SELECT statement. Does the solution meet the goal?

**No. You load the data using the CREATE TABLE AS SELECT statement**

[GE] You develop a data ingestion process that will import data to an enterprise data warehouse in Azure Synapse Analytics. The data to be ingested resides in parquet files stored in an Azure Data Lake Gen 2 storage account. You need to load the data from the Azure Data Lake Gen 2 storage account into the Data Warehouse.

Solution:

1. Create a remote service binding pointing to the Azure Data Lake Gen 2 storage account

2. Create an external file format and external table using the external data source

3. Load the data using the CREATE TABLE AS SELECT statement

Does the solution meet the goal?

**No. You need to create an external file format and external table from an external data source, instead from a remote service binding pointing.**

[GE] You develop a data ingestion process that will import data to a Microsoft Azure SQL Data Warehouse. The data to be ingested resides in parquet files stored in an Azure Data Lake Gen 2 storage account. You need to load the data from the Azure Data Lake Gen 2 storage account into the Azure SQL Data Warehouse.

Solution:

1. Create an external data source pointing to the Azure Data Lake Gen 2 storage account

2. Create an external file format and external table using the external data source

3. Load the data using the CREATE TABLE AS SELECT statement Does the solution meet the goal?

**Yes. You need to create an external file format and external table using the external data source. You load the data using the CREATE TABLE AS SELECT statement.**

<https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-load-from-azure-data-lake-store>

External data source

External file format

External table

Master key

[GE] Your company manages on-premises Microsoft SQL Server pipelines by using a custom solution. The data engineering team must implement a process to pull data from SQL Server and migrate it to Azure Blob storage. The process must orchestrate and manage the data lifecycle.  
You need to configure Azure Data Factory to connect to the on-premises SQL Server database.  
Which three actions should you perform in sequence?

[1] Create a virtual private network (VPN) connection from on-premises to Microsoft Azure. [2] Create an Azure Data Factory resource.

[3] Configure a self-hosted integration runtime.

Exam Prep:

Azure Data Factory

1. Understand the difference between all the available Integration runtime. Pay special attention to the self-hosted integration runtime.
2. Azure Data Factory Copy Activity: Find out schema mapping ways between source & sink. hands-on

# Lesson 10

INGESTION: Azure Synapse Analytics

**Azure Synapse Analytics** [Azure]

Azure Synapse Analytics is generalized analytics service. You can use it to read data from many sources, process this data, generate various analyses and models, and save the results. Azure Synapse Analytics uses a clustered architecture.

You can select between two technologies to process data:

* Transact-SQL
* Spark (same open-source technology used to power databricks)

**Synapse as a massively parallel processing database**

How to load data into Synapse Analytics: houseprices.csv

<https://docs.microsoft.com/en-us/learn/modules/explore-data-ingestion-azure/3-load-data>



1] Create Storage Account

*We use a storage account because ADLS is built on top of blob storage WASB*

storagedata

[2] Add File Share

housingdata

-- Upload csv file

houseprices.csv

[3] Create Synapse workspace Account

synapsews

-- Create Storage Account

synapsews-store

-- Add File Share

synapsews-data

[+] Create new SQL pool

synapsews-pool

Launch Synapse Studio

[+] Create new table

Housingdata

-- Publish table

Create new linked service (Source)

Create “Copy data” pipeline

(Destination)

[2] Create new Data Factory

synapsews-datafactory



*\*File share is like a replacement for a file server*

PREP & TRAIN: Azure Synapse Analytics

<https://docs.microsoft.com/en-us/azure/synapse-analytics/sql-data-warehouse/cheat-sheet>

Best Practice for building Azure Synapse Analytics Solutions **M.D.I.P.I**:

Diagram

Description automatically generated

|  |  |
| --- | --- |
| **Design** | **Recommendation** |
| Distribution | Round Robin |
| Indexing | Heap |
| Partition | None |
| Resource Class | largerc or xlargerc |

**Data migration**

First, load your data into Azure Data Lake Storage or Azure Blob Storage. Next, use the COPY statement (preview) to load your data into staging tables. Use the following configuration:

**Distributed or replicated tables**

Start with Round Robin, but aspire to a hash distribution strategy to take advantage of a massively parallel architecture. A distributed table appears as a single table, but the rows are stored across 60 distributions. The rows are distributed with a hash or round-robin algorithm.

|  |  |  |
| --- | --- | --- |
| **Type** | **Great fit for…** | **Watch out if** |
| Replicated | \* Small dimension tables in star schema (<2GB) | \* Many write transactions (insert, upsert, delete, update)  \* You change warehouse units (DWU) and provision frequently |
| Round Robin | \* Temp/Staging Table  No obvious joining key | \* Performance is slow due to data movement |
| Hash | \* Fact Tables  \* Large Dimension Tables | \* The distribution key cannot be updated |

**Index your tables**

Indexing is helpful for reading tables quickly. There is a unique set of technologies that you can use based on your needs:

|  |  |  |
| --- | --- | --- |
| **Type** | **Great fit for…** | **Watch out if** |
| Heap | Temp/Staging Table | \* Any lookup scans the full table |
| Clustered Index | Tables with up to 100 million rows  Large tables (more than 100 million rows) with only 1-2 columns heavily used | \* Used on a replicated table \* You have complex queries involving multiple join and Group By operations \* You make updates on the indexed columns: it takes memory |
| Clustered columnstore index (CGI) (default) | Large tables (more than 100 million rows) | \* Used on a replicated table  \* You make massive update operations on your table |

**Partitioning**

You might partition your table when you have a large fact table (greater than 1 billion rows). In 99 percent of cases, the partition key should be based on date. Be careful to not overpartition, especially when you have a clustered columnstore index.

With staging tables that require ELT, you can benefit from partitioning. It facilitates data lifecycle management. Be careful not to overpartition your data, especially on a clustered columnstore index.

**Incremental load**

If you're going to incrementally load your data, first make sure that you allocate larger resource classes to loading your data. This is particularly important when loading into tables with clustered columnstore indexes. See resource classes for further details.

We recommend using PolyBase and ADF V2 for automating your ELT pipelines into your data warehouse.

**Data Encryption**

How to secure sensitive data in a database in Azure SQL Database with data encryption by using the [Always Encrypted wizard](https://docs.microsoft.com/en-us/sql/relational-databases/security/encryption/always-encrypted-wizard) in [SQL Server Management Studio (SSMS)](https://docs.microsoft.com/en-us/sql/ssms/sql-server-management-studio-ssms).

* Create a column master key (CMK).
* Create a column encryption key (CEK).
* Create a database table and encrypt columns.
* Create an application that inserts, selects, and displays data from the encrypted columns.

*MCED — Master Key, Certificate, Encryption & Apply encryption on the DB*

NB: Use the Azure Key Vault, not the Windows Certificate Store, to store the master key. Use the Always Encrypted wizard in SSMS to create Always Encrypted keys.

Data Security

Azure Storage encrypts all data that's written to it. Azure Storage also provides you with fine-grained control over who has access to your data. You'll secure the data by using keys or shared access signatures.

Azure Resource Manager provides a permissions model that uses role-based access control (**RBAC**). Use this functionality to set permissions and assign roles to users, groups, or applications.

[GE] You have an Azure SQL database named DB1 that contains a table named Table1. Table1 has a field named Customer\_ID that is varchar(22). You need to implement masking for the Customer\_ID field to meet the following requirements:

- The first two prefix characters must be exposed.

- The last four prefix characters must be exposed.

- All other characters must be masked.

Solution: You implement data masking and use a random number function mask. Does this meet the goal?

**No. Must use Custom Text data masking, which exposes the first and last characters and adds a custom padding string in the middle**

Solution: You implement data masking and use an email function mask. Does this meet the goal?

**No. Must use Custom Text data masking, which exposes the first and last characters and adds a custom padding string in the middle**

Solution: You implement data masking and use a credit card function mask. Does this meet the goal?

**No. Must use Custom Text data masking, which exposes the first and last characters and adds a custom padding string in the middle**

Solution: You implement data masking and use a custom text mask. Does this meet the goal?

**Yes. We must use Custom Text data masking, which exposes the first and last characters and adds a custom padding string in the middle.**

[GE] You have an Azure SQL database that has masked columns. You need to identify when a user attempts to infer data from the masked columns. What should you use?

A. Azure Advanced Threat Protection (ATP)

B. custom masking rules

C. Transparent Data Encryption (TDE)

**D. auditing**

Dynamic Data Masking is designed to simplify application development by limiting data exposure in a set of pre-defined queries used by the application. While Dynamic Data Masking can also be useful to prevent accidental exposure of sensitive data when accessing a production database directly, it is important to note that unprivileged users with ad-hoc query permissions can apply techniques to gain access to the actual data. If there is a need to grant such ad-hoc access, Auditing should be used to monitor all database activity and mitigate this scenario.

[GE] Your company manages a payroll application for its customers worldwide. The application uses an Azure SQL database named DB1. The database contains a table named Employee and an identity column named EmployeeId. A customer requests the EmployeeId be treated as sensitive data. Whenever a user queries EmployeeId, you need to return a random value between 1 and 10 instead of the EmployeeId value. Which masking format should you use?

A. string

**B. number**

C. default

<https://docs.microsoft.com/en-us/azure/sql-database/sql-database-dynamic-data-masking-get-started-portal>

[GE] You develop data engineering solutions for a company. A project requires the deployment of data to Azure Data Lake Storage. You need to implement role-based access control (RBAC) so that project members can manage the Azure Data Lake Storage resources. Which three actions should you perform?

**A. Assign Azure AD security groups to Azure Data Lake Storage.**

B. Configure end-user authentication for the Azure Data Lake Storage account.

C. Configure service-to-service authentication for the Azure Data Lake Storage account.

**D. Create security groups in Azure Active Directory (Azure AD) and add project members.**

**E. Configure access control lists (ACL) for the Azure Data Lake Storage account.**

[GE] A company runs Microsoft SQL Server in an on-premises virtual machine (VM). You must migrate the database to Azure SQL Database. You synchronize users from Active Directory to Azure Active Directory (Azure AD). You need to configure Azure SQL Database to use an Azure AD user as administrator. What should you configure?

A. For each Azure SQL Database, set the Access Control to administrator.

B. For each Azure SQL Database server, set the Active Directory to administrator.

**C. For each Azure SQL Database, set the Active Directory administrator role.**

D. For each Azure SQL Database server, set the Access Control to administrator.

There are two administrative accounts (Server admin and Active Directory admin) that act as administrators. One Azure Active Directory account, either an individual or security group account, can also be configured as an administrator. It is optional to configure an AzureAD administrator, but an Azure AD administrator must be configured if you want to use Azure AD accounts to connect to SQL Database.

<https://docs.microsoft.com/en-us/azure/sql-database/sql-database-manage-logins>

Create security groups in Azure Active Directory. Assign users or security groups to Data Lake Storage Gen1 account.

Assign users or security groups as ACLs to the Data Lake Storage Gen1 file system

[GE] You have an Azure SQL database that contains a table named Customer. Customer contains the columns shown in the following table.

|  |  |  |
| --- | --- | --- |
| Customer\_ID | Customer\_name | Customer\_Email |
| 44531 | John Smith | johns@contoso.com |
| 44532 | Tom Jones | tomj@nwtraders.com |
| 44533 | Bill Taylor | billit@litwareinc.com |

You apply a masking rule as shown in the following table. Which users can view the email addresses of the customers?

|  |  |  |  |
| --- | --- | --- | --- |
| Mask name | Target column | Mask function | User excluded |
| Customer Emails | Customer\_Email | Email | None |

A. Server administrators and all users who are granted the UNMASK permission to the Customer\_Email column only.

**B. All users who are granted the UNMASK permission to the Customer\_Email column only.**

C. Server administrators only.

D. Server administrators and all users who are granted the SELECT permission to the Customer\_Email column only.

*Server administrators do not have automatic access to masked columns!*

Grant the UNMASK permission to a user to enable them to retrieve unmasked data from the columns for which masking is defined.

<https://docs.microsoft.com/en-us/sql/relational-databases/security/dynamic-data-masking>

[GE] On which data store you configure Transparent data encryption (TDE) to meet the technical requirements?

A. Cosmos DB

**B. Azure Synapse Analytics**

C. SQL Database

Scenario: Transparent data encryption (TDE) must be enabled on all data stores, whenever possible. The database for Mechanical Workflow must be moved to Azure Synapse Analytics.

*Cosmos DB does not support TDE!*

[GE] You develop data engineering solutions for a company. You must migrate data from Microsoft Azure Blob storage to an Azure SQL Data Warehouse for further transformation. You need to implement the solution. Which four actions should you perform in sequence?

**Step 1:** Provision an Azure SQL Data Warehouse instance.  
Create a data warehouse in the Azure portal.  
**Step 2:** Connect to the Azure SQL Data warehouse by using SQL Server Management Studio  
Connect to the data warehouse with SSMS (SQL Server Management Studio)  
**Step 3:** Build external tables by using the SQL Server Management Studio  
Create external tables for data in Azure blob storage.  
You are ready to begin the process of loading data into your new data warehouse. You use external tables to load data from the Azure storage blob.  
**Step 4:** Run Transact-SQL statements to load data.

[GE] You need to configure data encryption for external applications.

Solution:

1. Access the Always Encrypted Wizard in SQL Server Management Studio

2. Select the column to be encrypted

3. Set the encryption type to Deterministic

4. Configure the master key to use the Windows Certificate Store

5. Validate configuration results and deploy the solution

Does the solution meet the goal?

**No.**

**Polybase**

Feature of SQL Server and Azure Synapse Analytics

Enables you to run Transact-SQL queries that read data from external data sources (makes these sources appear like SQL tables).

Data Factory can directly invoke PolyBase on your behalf if your data is in a PolyBase-compatible data store.

Polybase

[GE] You have an Azure SQL data warehouse. Using PolyBase, you create table named [Ext].[Items] to query Parquet files stored in Azure Data Lake Storage Gen2 without importing the data to the data warehouse. The external table has three columns. You discover that the Parquet files have a fourth column named ItemID. Which command should you run to add the ItemID column to the external table?

Text

Description automatically generated

**A.Option A**

B.Option B

C.Option C

D.Option D

[GE] You have an enterprise data warehouse in Azure Synapse Analytics. You need to monitor the data warehouse to identify whether you must scale up to a higher service level to accommodate the current workloads. Which is the best metric to monitor? More than one answer choice may achieve the goal. Select the BEST answer.

A.CPU percentage

**B.DWU used**

C.DWU percentage

D.Data IO percentage

DWU used, defined as DWU limit \* DWU percentage, represents only a high-level representation of usage across the SQL pool and is not meant to be a comprehensive indicator of utilization. To determine whether to scale up or down, consider all factors which can be impacted by DWU such as concurrency, memory, tempdb, and adaptive cache capacity. We recommend running your workload at different DWU settings to determine what works best to meet your business objectives.

<https://docs.microsoft.com/bs-latn-ba/azure/synapse-analytics/sql-data-warehouse/sql-data-warehouse-concept-resource-utilization-query-activity>

[GE] You configure monitoring for a Microsoft Azure SQL Data Warehouse implementation. The implementation uses PolyBase to load data from comma-separated value (CSV) files stored in Azure Data Lake Gen 2 using an external table. Files with an invalid schema cause errors to occur. You need to monitor for an invalid schema error. For which error should you monitor?

*PolyBase is recursively reading each file in the directory (CREATE EXTERNAL TABLE DDL points to this directory), where each file has a different schema.*

A. EXTERNAL TABLE access failed due to internal error:

'Java exception raised on call to HdfsBridge\_Connect: Error[com.microsoft.polybase.client.KerberosSecureLogin] occurred while accessing external file.'

*Kerberos is not enabled in Hadoop Cluster*

B. EXTERNAL TABLE access failed due to internal error:

'Java exception raised on call to HdfsBridge\_Connect:

Error [NoFileSystem for scheme: wasbs] occurred while accessing external file.'

*wasbs refers to windows azure storage blob*

**C. Cannot execute the query "Remote Query" against OLE DB provider "SQLNCLI11":**

**for linked server "(null)",**

**Queryaborted- the maximum reject threshold (o rows) was reached while reading from an external source: 1 rows rejected out of total 1 rows processed.**

*This error would be seen in SSMS (column ot data type mismatch).*

D. EXTERNAL TABLE access failed due to internal error:

'Java exception raised on call to HdfsBridge\_Connect:

Error[Unable to instantiate LoginClass] occurred while accessing external file.'

Possible Solution: If the data for each table consists of one file, then use the filename in the LOCATION section prepended by the directory of the external files. If there are multiple files per table, put each set of files into different directories in Azure Blob Storage and then you can point LOCATION to the directory instead of a particular file. suggestion is the best practices recommended by SQLCAT even if you have one file per table.

[GE] You are developing a solution to visualize multiple terabytes of geospatial data.

The solution has the following requirements:

- Data must be encrypted.

- Data must be accessible by multiple resources on Microsoft Azure.

You need to provision storage for the solution.

Which four actions should you perform in sequence?

**[1] Create a new Azure Data Lake Storage account with Azure Key Vault managed encryption keys**

**[2] Select and configure an encryption key storage container.**

**[3] Add an access policy for the new Azure Data Lake account to the key storage container.**

**[4] Enable encryption on the Azure Data Lake using the Azure portal**

[GE] You are responsible for providing access to an Azure Data Lake Storage Gen2 account. Your user account has contributor access to the storage account, and you have the application ID access key. You plan to use PolyBase to load data into Azure SQL data warehouse. You need to configure PolyBase to connect the data warehouse to the storage account.

Which three components should you create in sequence?

**[1] a database scoped credential**

*need to create a Database Master Key to encrypt your credential secret used in the next step.*

**[2] an external data source**

*Use the CREATE EXTERNAL DATA SOURCE command to store the location of the data*

**[3] an external file format**

*To import the data from Data Lake Storage, you need to specify the External File Format*

[GE] You manage the Microsoft Azure Databricks environment for a company. You must be able to access a private Azure Blob Storage account. Data must be available to all Azure Databricks workspaces. You need to provide the data access.  
Which three actions should you perform in sequence?

|  |  |
| --- | --- |
| Upload a certificate |  |
| Add secrets to the scope |  |
| Use Blob storage access key |  |
| Create a secret scope |  |
| Configure a JDBC connector |  |
| Mount the Azure Blob Storage container |  |

**[1]** **Create a secret scope**

**[2] Add secrets to the scope**

**[3]** **Mount the Azure Blob Storage container**

[GE] A company uses Microsoft Azure SQL Database to store sensitive company data. You encrypt the data and only allow access to specified users from specified locations.

You must monitor data usage, and data copied from the system to prevent data leakage.

You need to configure Azure SQL Database to email a specific user when data leakage occurs.

Which three actions should you perform in sequence?

|  |  |
| --- | --- |
| In Auditing, enable Auditing |  |
| Configure the service to create alters for threat detections of type Data Exfiltration. |  |
| In Firewalls and virtual networks, enable Allow access to Azure services. |  |
| Enable advanced threat protection. |  |
| Configure the service to send email alerts to security@contoso.com |  |

**[1]** **Enable advanced threat protection**

**[2] Configure the service to send email alerts to security@contoso.team**

[3] Configure the service to create alters for threat detections of type Data Exfiltration.

Exam prep:

Azure SQL Server & SQL Server Data Warehouse (Synapse Analytics)

1. Which SQL option should I choose?
2. Export an Azure SQL database to a BACPAC file. hands-on
3. Learn about how to secure sensitive data in a SQL database with database encryption by using the Always Encrypted wizard. hands-on
4. Experience Azure SQL Database Advanced Threat Protection features & steps to enable it. hands-on
5. Experiment and enable TDE (Transparent Data Encryption) and keep a note on the steps: Formula(memory trick): MCED — Master Key, Certificate, Encryption & Apply encryption on the DB. hands-on
6. Do an experiment using Powershell & Azure cloud shell. hands-on
7. IP firewall rules. hands-on
8. Read about dynamic data masking for Azure SQL Database and Azure Synapse Analytics. Give special attention to the in-built masking functions & their appropriate usages (Default, Credit Card, Email, Random Number, Custom Text)
9. Polybase: Please execute this hands-on experiment multiple times to load the data from ADLS into WH and memorize all the steps in the correct sequence. Formula(Memory trick): MCSFTL — Master, Credential, Source, File, Table, Load(CTAS). Load New York Taxicab dataset hands-on
10. DW performance benchmarking: This example demonstrates DW performance benchmarking and concluded to utilize a methodology of CTAS and partition switching in lieu of UPDATE and DELETE operations wherever possible. Get a full understanding of this fundamental approach. hands-on

# Lesson 11

INGESTION: Azure Databricks

**Azure Databricks** [Azure]



Azure Databricks is an analytics platform optimized for the Microsoft Azure cloud services platform. Databricks is based on Spark, and is integrated with Azure to streamline workflows.

It provides an interactive workspace that enables collaboration between data scientists, data engineers, and business analysts.

Databricks can process data held in many different types of storage, including Azure Blob storage, Azure Data Lake Store, Hadoop storage, flat files, SQL databases, and data warehouses, and Azure services such as Cosmos DB. Databricks can also process streaming data

**Cluster configurations**

Databricks supports three cluster modes: Standard, High Concurrency, and Single Node.

|  |  |  |
| --- | --- | --- |
| **Standard** | **High Concurrency** | **Single Node** |
| Single User | Multiple Users |  |
| Terminates automatically after 120 minutes | Does not terminate automatically | Terminates automatically after 120 minutes |
|  |  |  |
| Any language - Python, R, Scala, and SQL | SQL, Python and R |  |
|  | Notebook Isolation |  |

*You cannot change the cluster mode after a cluster is created. If you want a different cluster mode, you must create a new cluster*

**Cluster types**

Databricks supports two types of clusters: interactive and automated

|  |  |
| --- | --- |
| **Interactive** | **Automated** |
| use interactive clusters to analyze data collaboratively with interactive notebooks | automated clusters to run fast and robust automated jobs. |
|  | Scheduled batch workloads (data engineers running ETL jobs) |

[GE] You plan to perform batch processing in Azure Databricks once daily. Which type of Databricks cluster should you use?

**A. automated**

B. interactive

C. High Concurrency

Azure Databricks has two types of clusters: interactive and automated. You use interactive clusters to analyze data collaboratively with interactive notebooks. You use automated clusters to run fast and robust automated jobs. Example: Scheduled batch workloads (data engineers running ETL jobs) This scenario involves running batch job JARs and notebooks on a regular cadence through the Databricks platform. The suggested best practice is to launch a new cluster for each run of critical jobs. This helps avoid any issues (failures, missing SLA, and so on) due to an existing workload (noisy neighbor) on a shared cluster.

<https://docs.databricks.com/administration-guide/cloud-configurations/aws/cmbp.html#scenario-3-scheduled-batch-workloads-data-engineers-running-etl-jobs>

[GE] You plan to create an Azure Databricks workspace that has a tiered structure. The workspace will contain the following three workloads:

- A workload for data engineers who will use Python and SQLA workload for jobs that will run notebooks that use Python, Spark, Scala, and

- SQLA workload that data scientists will use to perform ad hoc analysis in Scala and R

The enterprise architecture team at your company identifies the following standards for Databricks environments:

- The data engineers must share a cluster.

- The job cluster will be managed by using a request process whereby data scientists and data engineers provide packaged notebooks for deployment to the cluster.

- All the data scientists must be assigned their own cluster that terminates automatically after 120 minutes of inactivity. Currently, there are three data scientists.

- You need to create the Databrick clusters for the workloads.

Solution: You create a Standard cluster for each data scientist, a High Concurrency cluster for the data engineers, and a High Concurrency cluster for the jobs.

Does this meet the goal?

**Yes. A high concurrency cluster is a managed cloud resource. The key benefits of high concurrency clusters are that they provide Apache Spark-native fine-grained sharing for maximum resource utilization and minimum query latencies.**

Solution: You create a Standard cluster for each data scientist, a High Concurrency cluster for the data engineers, and a Standard cluster for the jobs. Does this meet the goal?

**No. We would need a High Concurrency cluster for the jobs.**

[GE] You are creating a new notebook in Azure Databricks that will support R as the primary language but will also support Scola and SQL. Which switch should you use to switch between languages?

**A. %<language>**

B. \\[<language>]

C. \\(<language>)

D. @<Language>

You can override the primary language by specifying the language magic command %<language> at the beginning of a cell. The supported magic commands are: %python, %r, %scala, and %sql

[GE] You plan to perform batch processing in Azure Databricks once daily. Which type of Databricks cluster should you use?

**A. job**

B. interactive

C. High Concurrency

Example: Scheduled batch workloads (data engineers running ETL jobs) This scenario involves running batch job JARs and notebooks on a regular cadence through the Databricks platform. The suggested best practice is to launch a new cluster for each run of critical jobs. This helps avoid any issues (failures, missing SLA, and so on) due to an existing workload (noisy neighbor) on a shared cluster.

Note: Azure Databricks has two types of clusters: interactive and automated. You use interactive clusters to analyze data collaboratively with interactive notebooks.

You use automated clusters to run fast and robust automated jobs.

<https://docs.databricks.com/administration-guide/cloud-configurations/aws/cmbp.html#scenario-3-scheduled-batch-workloads-data-engineers-running-etl-jobs>

Exam prep:

Azure Databricks

1. Learn about the technology choices for batch processing and what is the decision criteria to choose one over the others.
2. ETL using Azure Databricks. Special attention to “Load data into Azure SQL Data Warehouse” hands-on
3. Experiment on different cluster configurations. hands-on

# Lesson 12

Hadoop

A picture containing drawing

Description automatically generatedHadoop was developed to solve the following:

|  |  |
| --- | --- |
| **Challenges** | **Solutions** |
| Single central storage | Distributed storage |
| Serial (linear) processing | Map reduce: Parallel processing |
| Lack of ability to process unstructured data | Ability to process every type of data |

Hadoop is an open-source framework that manages big data storage in a distributed way and processes it parallelly. It does this processing and analysis of big data on clusters. The Hadoop technology stack includes related software and utilities including Apache Hive, Apache HBase, Spark, Kafka etc.

Apache Hadoop framework:

**Hadoop Common**

**Hadoop Distributed File System (HDFS)**

**Hadoop YARN**

**Hadoop MapReduce**

Diagram

Description automatically generated

A picture containing drawing

Description automatically generated

Storage unit

HDFS : Specially designed for storing huge datasets in commodity hardware

Master/ name node

-- slave/ data node

-- slave/ data node

-- slave/ data node

Processing unit

Map Reduce: is a programming technique where huge data is processed in a parallel and distributed fashion (data is processed at slave nodes)

Master/ name node

-- slave/ data node

-- slave/ data node

-- slave/ data node

How to install Hadoop (Pseudo-Distributed Mode)

*Practically, you would use something like docker to set this up*

<https://medium.com/@thedsa.in/install-hadoop-3-2-setting-up-a-single-node-hadoop-cluster-22a5754bd9fc>

[1] Create CentOS VM

-- Create Hadoop user

adduser hduser

passwd hduser

usermod -aG wheel hduser

[2] Install Java (CentOS)

<https://www.liquidweb.com/kb/install-java-8-on-centos-7/>

yum -y update

yum install java-1.8.0-openjdk

yum install java-1.8.0-openjdk-headless

[3] Setup SSH

Install OpenSSH Server

sudo yum –y install openssh-server openssh-clients

# start SSH daemon on the openSSH server

sudo systemctl start sshd

sudo systemctl status sshd

[3] Configure SSH key-based Authentication

su – hduser

ssh-keygen -t rsa

cat ~/.ssh/id\_rsa.pub >> ~/.ssh/authorized\_keys

chmod 640 ~/.ssh/authorized\_keys

# test user

ssh localhost

[4] Download and Configure Hadoop

wget http://apachemirror.wuchna.com/hadoop/common/hadoop-3.2.1/hadoop-3.2.1.tar.gz

tar -xvzf hadoop-3.2.1.tar.gz

mv hadoop-3.2.1 hadoop

[5] Configure .bashrc and env variables

update-alternatives --config java

# copy the above path

vim .bashrc

#paste above path

export JAVA\_HOME=/usr/lib/jvm/java-1.8.0-openjdk-1.8.0.191.b12-1.el7\_6.x86\_64/jre/

export HADOOP\_HOME=/home/hduser/hadoop

export HADOOP\_INSTALL=$HADOOP\_HOME

export HADOOP\_MAPRED\_HOME=$HADOOP\_HOME

export HADOOP\_COMMON\_HOME=$HADOOP\_HOME

export HADOOP\_HDFS\_HOME=$HADOOP\_HOME

export HADOOP\_YARN\_HOME=$HADOOP\_HOME

export HADOOP\_COMMON\_LIB\_NATIVE\_DIR=$HADOOP\_HOME/lib/native

export PATH=$PATH:$HADOOP\_HOME/sbin:$HADOOP\_HOME/bin

export HADOOP\_OPTS="-Djava.library.path=$HADOOP\_HOME/lib/native"

source ~/.bashrc

echo $JAVA\_HOME

[6] Configure Hadoop

mkdir -p ~/hadoopdata/hdfs/namenode

mkdir -p ~/hadoopdata/hdfs/datanode

vim $HADOOP\_HOME/etc/hadoop/core-site.xml

vim $HADOOP\_HOME/etc/hadoop/hdfs-site.xml

vim $HADOOP\_HOME/etc/hadoop/mapred-site.xml

vim $HADOOP\_HOME/etc/hadoop/yarn-site.xml

[7] Start Hadoop Cluster

hdfs namenode -format

# the above will give you the hostname (VM name)

# SHUTDOWN\_MSG: Shutting down NameNode at AN-01/45.58.38.202

start-dfs.sh

start-yarn.sh

# run the jps command. You should see 6 services running

jps

[7] Configure Firewall

# OPTIONAL: Enable Firewalld

systemctl enable firewalld

systemctl start firewalld

systemctl status firewalld

# allow connections through firewall

sudo su

firewall-cmd --permanent --add-port=9870/tcp

firewall-cmd --permanent --add-port=8088/tcp

firewall-cmd --reload

Forward ports (~/.ssh/config)

Host AN\*

User hduser

Compression yes

ForwardAgent yes

Host AN-01

HostName 102.37.120.33

# NameNode

LocalForward 0.0.0.0:9870 localhost:9870

# ResourceManager

LocalForward 0.0.0.0:8088 localhost:8088

# NodeManager

LocalForward 0.0.0.0:8042 localhost:8042

# DataNode

LocalForward 0.0.0.0:9864 slocalhost:9864

# DataNode

LocalForward 0.0.0.0:9866 slocalhost:9866

[GE] You are a data engineer. You are designing a Hadoop Distributed File System (HDFS) architecture. You plan to use Microsoft Azure Data Lake as a data storage repository. You must provision the repository with a resilient data schema. You need to ensure the resiliency of the Azure Data Lake Storage. What should you use? To answer, select the appropriate options in the answer area.

**NameNode**

An HDFS cluster consists of a single NameNode, a master server that manages the file system namespace and regulates access to files by clients.

**DataNode**

The DataNodes are responsible for serving read and write requests from the file system's clients.

**DataNode**

The DataNodes perform block creation, deletion, and replication upon instruction from the NameNode.

Note: HDFS has a master/slave architecture. An HDFS cluster consists of a single NameNode, a master server that manages the file system namespace and regulates ac

# Lesson 13

INGESTION + STORAGE: Azure Synapse Analytics

**Azure HDInsight** [Azure]



Azure HDInsight is a managed analytics service in the cloud. It is basically an implementation of Hadoop in Azure.

HDInsight is a low-cost cloud solution. It includes Apache Hadoop, Spark, Kafka, HBase, Storm, and Interactive Query.

These enable you to run processing tasks over large amounts of data

HDInsight uses a clustered model, like that of Synapse Analytics. HDInsight stores data using Azure Data Lake storage. Hadoop Map/Reduce uses a simple framework to split a task over a large dataset into a series of smaller tasks over subsets of the data that can be run in parallel, and the results then combined.

**To Query Hadoop supports Pig and HiveQL languages. In Spark, data engineers use Spark SQL.**

HDInsight Cluster types

**Apache Hadoop:** A framework that uses HDFS (storage), YARN (resource management), and a simple MapReduce programming model to process and analyze batch data in parallel.

**Apache Spark:** A parallel processing framework that supports in-memory processing to boost the performance of big-data analysis applications. Spark works for SQL, streaming data, and machine learning   
*This makes Spark about 100 times faster*

**Apache HBase:** A NoSQL database built on Hadoop that provides random access and strong consistency for large amounts of unstructured and semi-structured data–potentially billions of rows times millions of columns

*It's commonly used for search engines, and has automatic failover*

**Microsoft R Server:** A server for hosting and managing parallel, distributed R processes. It provides data scientists, statisticians, and R programmers with on-demand access to scalable, distributed methods of analytics on HDInsight.

**Apache Storm:** A distributed, real-time computation system for processing large streams of data fast. Storm is offered as a managed cluster in HDInsight

**Apache Kafka:** An open-source platform that’s used for building streaming data pipelines and applications. Kafka also provides message-queue functionality that allows you to publish and subscribe to data streams.

**Apache Interactive** In-memory caching for interactive and faster Hive queries

**Query preview**

**Apache Hive**

**LLAP**

|  |  |
| --- | --- |
| HDInsight Cluster Type | Workload |
| Hadoop | ETL/ELT |
| Storm | Data in Motion / IoT |
| HBase | Transactional Processing |
| Spark -or- R Server with Spark | Data Science / Advanced Analytics |

How to create data lake storage:

1

[1] Create hivescript.hql

Logo

Description automatically generated

DROP TABLE IF EXISTS HiveSampleOut;

CREATE EXTERNAL TABLE HiveSampleOut (clientid string, market string, devicemodel string, state string)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ' '

STORED AS TEXTFILE LOCATION 'S{hiveconf:output}';

INSERT OVERWRITE TABLE HiveSampleOut

SELECT

clientid,

market,

devicemodel,

state

FROM hivesampletable

[2] Create Storage Account

*Not gen 2 so no hierarchical storage needed*

hdinsighthive

-- Add Container (blob)

hdinsight

-- Upload hivescript.hql to blob

[6] Create new linked service

Manage > Linked Service > Data Store

**Azure Blob**

[7] Create new linked service

Manage > Linked Service > Compute

**HDInsight**

*Copy your principal ID and key into here*

-- Azure Storage linked service [6]

-- Cluster Type: Hadoop

-- Cluster Size: Hadoop

*Create Hadoop user*

-- OS Type:

ssh username: hduser

ssh password: <passowrd>

[3] Create App Registration (AAD)  
hdinsighthive.com

-- Create client secret

-- Copy App ID

[4] Add Role to Resource Group  
*Add contributor role to hdinsighthive service principal*

Resource Group > IAM > Add Role Assignment

[5] Create new Data Factory

hivefactory

Icon

Description automatically generated

[8] Create data Pipeline

-- Create “HDInsight” Hive activity

-- HDI Cluser: linkedservice1

-- Script: hdinsight/hivescript.hpl

-- Parameters: auto-fill from script

Output: DRIVER\_NAME://CONTAINER\_NAME@STORAGE\_ACCOUNT.blob.core.windows.net/outputfolder

wasbs://hdinsight@hdinsighthive.blob.core.windows.net/outputfolder

[GE] You are a data engineer implementing a lambda architecture on Microsoft Azure. You use an open-source big data solution to collect, process, and maintain data. The analytical data store performs poorly. You must implement a solution that meets the following requirements:

- Provide data warehousing

- Reduce ongoing management activities

- Deliver SQL query responses in less than one second

You need to create an HDInsight cluster to meet the requirements. Which type of cluster should you create? **Apache Spark**

[GE] You need to develop a pipeline for processing data. The pipeline must meet the following requirements:

- Scale up and down resources for cost reduction

- Use an in-memory data processing engine to speed up ETL and machine learning operations.

- Use streaming capabilities

- Provide the ability to code in SQL, Python, Scala, and R

- Integrate workspace collaboration with Git What should you use?

**A. HDInsight Spark Cluster**

B. Azure Stream Analytics

C. HDInsight Hadoop Cluster

D. Azure SQL Data Warehouse

E. HDInsight Kafka Cluster

F. HDInsight Storm Cluster

[GE] The data engineering team manages Azure HDInsight clusters. The team spends a large amount of time creating and destroying clusters daily because most of the data pipeline process runs in minutes.

You need to implement a solution that deploys multiple HDInsight clusters with minimal effort.

What should you implement?

A. Azure Databricks

B. Azure Traffic Manager

**C. Azure Resource Manager templates**

D. Ambari web user interface

[GE] Your company plans to create an event processing engine to handle streaming data from Twitter. The data engineering team uses Azure Event Hubs to ingest the streaming data.

You need to implement a solution that uses Azure Databricks to receive the streaming data from the Azure Event Hubs.

Which three actions should you recommend be performed in sequence?

**[1] Deploy the Azure Databricks service**

**[2] Deploy a Spark cluster and then attach the required libraries to the cluster.**

**[3] Create and configure a Notebook that consumes the streaming data.**

[GE] You are developing a solution using a Lambda architecture on Microsoft Azure. The data at rest layer must meet the following requirements:

Data storage:

- Serve as a repository for high volumes of large files in various formats.

- Implement optimized storage for big data analytics workloads.

- Ensure that data can be organized using a hierarchical structure.

Batch processing:

- Use a managed solution for in-memory computation processing.

- Natively support Scala, Python, and R programming languages.

Provide the ability to resize and terminate the cluster automatically.

Analytical data store:

- Support parallel processing.

- Use columnar storage.

- Support SQL-based languages.

You need to identify the correct technologies to build the Lambda architecture.

Which technologies should you use?

**Data Storage: Azure Data Lake Store**

*Allows hierarchical data store*

**Batch Processing: HDInsight Spark**

*Open-source, parallel-processing framework that supports in-memory processing*

**Analytical data store: Azure SQL Data Warehouse**

*SQL Data Warehouse stores data into relational tables with columnar storage.*

[GE] You are designing a new Lambda architecture on Microsoft Azure.

The real-time processing layer must meet the following requirements:

Ingestion:

- Receive millions of events per second

- Act as a fully managed Platform-as-a-Service (PaaS) solution

- Integrate with Azure Functions

Stream processing:

- Process on a per-job basis

- Provide seamless connectivity with Azure services

- Use a SQL-based query language

Analytical data store:

- Act as a managed service

- Use a document store

- Provide data encryption at rest

You need to identify the correct technologies to build the Lambda architecture using minimal effort. Which technologies should you use?

**Data Storage: Azure Event Hubs**

*Options include Azure Event Hubs, Azure IoT Hub, and Kafka*

**Batch Processing: Azure Stream Analytics**

*based on perpetually running SQL queries that operate on unbounded streams.*

**Analytical data store: Azure SQL Data Warehouse**

*Azure SQL Data Warehouse provides a managed service for large-scale, cloud-based data warehousing*

Storage & HDInsight

1. Get in-depth knowledge of using Azure Data Lake Storage Gen2 for big data requirements. Also, learn about different Hadoop tools discussed in this article.
2. Choose the correct HDInsight Configuration to build open-source analytics solutions. Give special attention to the use cases and get a better understanding of when to use Storm vs Spark etc.

# Lesson 14

STORAGE: Azure Data Lake Storage Gen 2

<https://www.dremio.com/data-lake/azure/>

**Azure Data Lake Storage Gen 2** [Azure]

Icon

Description automatically generatedis a fully managed, elastic, scalable and secure file system that supports HDFS semantics and works with the Apache Hadoop ecosystem. ADLS can store unstructured, semi-structed and structured data. It has no upfront cost (pay-per-use model).

it is a combination of Azure Blob Storage + Azure Data Lake Gen 1.

**Gen 1 Features:**

Azure Data Lake Storage is a Hadoop-compatible data repository that can store any

size or type of data.

**Gen 2 Features:**

- Unlimited scalability

- Cost effective (Pay-per-use model)

- Hadoop compatibility (Manage and access data the same way you would with HDFS)

- Geo-redundant storage

- Zone-redundant storage

- Security support for both access control lists **ACL’s**

- POSIX compliance

- An optimized Azure Blob File System **ABFS** driver (optimized for big data analytics)

Diagram, funnel chart

Description automatically generated

Exam Prep:

Storage

1. Manage the Azure Blob storage lifecycle.
2. Experiment the mentioned example in this post about applying a lifecycle policy. hands-on
3. Learn about Access control in Azure Data Lake Storage Gen2. Pay special
4. attention to Azure AD setup while applying ACLs. hands-on
5. Configure Azure Storage firewalls and virtual networks. hands-on

How to create data lake storage:

1

[1] Create Storage Account

*We use a storage account because ADLS is built on top of blob storage WASB*

adlstorage

-- enable Hierarchical namespace (gen 2)

**Microsoft Azure Storage explorer (Windows/Mac)**[2] Add Container (blob)

images

-- Add new folder

personal

-- Upload file (random local file)

personal

Icon

Description automatically generated

Icon

Description automatically generated

[3] Create new Data Factory

adlsfactory

A picture containing icon

Description automatically generated

[4] Create new SQL Database

adlsdb

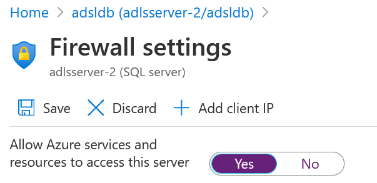
-- Add server

adlsserver

NB: remember to toggle [Azure]

“Allow Azure services and resources to access this server”

-- Add admin user to server



# Lesson 15

PREP & TRAIN: Azure Stream Analytics

[Azure]

<https://docs.microsoft.com/en-us/learn/modules/survey-the-azure-data-platform/8-stream-analytics>

Data engineers use Azure Stream Analytics to process streaming data and respond to data anomalies in real time. You can use Stream Analytics for Internet of Things (IoT) monitoring, web logs, remote patient monitoring, and point of sale (POS) systems.

**Use the declarative Stream Analytics query language to query Azure Stream Analytics**

*Diagram

Description automatically generated*

Stream Analytics

A Stream Analytics job definition includes at least one:  
**Input (where the job reads data stream from)**

EventHub (partition key explicitly with PARTITION BY keyword, compatibility level 1.1)

IoT Hub (partition key explicitly with PARTITION BY keyword, compatibility level 1.1)

Blob storage

**Query (transform the data input stream)**

**Output (where the job sends the results to)**

Azure Data Lake Storage

Azure Functions

Azure Table

Blob storage (can set the partition key explicitly)

Cosmos DB (need to set the partition key explicitly)

Event Hubs (need to set the partition key explicitly)

IoT Hub (need to set the partition key explicitly)

Service Bus

SQL and Azure Synapse Analytics with optional partitioning

**Stream Analytics: Query parallelization**

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization>

Partitioning lets you divide data into subsets based on a partition key. If your input (for example Event Hubs) is partitioned by a key, it is highly recommended to specify this partition key when adding input to your Stream Analytics job. Scaling a Stream Analytics job takes advantage of partitions in the input and output. A Stream Analytics job can consume and write different partitions in parallel, which increases throughput

Embarrassingly parallel

In parallel computing, an embarrassingly parallel workload, or embarrassingly parallel problem, is one for which little or no effort is required to separate the problem into several parallel tasks.

**Stream Analytics windowing functions [T.H.S.S.S]**

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions>

Timeline

Description automatically generatedIn time-streaming scenarios, performing operations on the data contained in temporal windows is a common pattern. Stream Analytics has native support for windowing functions, enabling developers to author complex stream processing jobs with minimal effort. Tumbling, Hopping, Sliding, Session, and Snapshot windows

Tumbling window functions are used to segment a data stream into distinct time segments and perform a function against them, such as the example below. The key differentiators of a Tumbling window are that they repeat, do not overlap, and an event cannot belong to more than one tumbling window.

Timeline

Description automatically generatedHopping window functions hop forward in time by a fixed period. It may be easy to think of them as Tumbling windows that can overlap and be emitted more often than the window size. Events can belong to more than one Hopping window result set. To make a Hopping window the same as a Tumbling window, specify the hop size to be the same as the window size.

Chart, timeline

Description automatically generatedSliding windows, unlike Tumbling or Hopping windows, output events only for points in time when the content of the window changes. In other words when an event enters or exits the window. So, every window has at least one event. Like Hopping windows, events can belong to more than one sliding window.

Timeline

Description automatically generatedSession window functions group events that arrive at similar times, filtering out periods of time where there is no data. It has three main parameters: timeout, maximum duration, and partitioning key (optional).

Graphical user interface, application

Description automatically generatedSnapshot windows groups events that have the same timestamp. Unlike other windowing types, which require a specific window function (such as [SessionWindow()](https://docs.microsoft.com/en-us/stream-analytics-query/session-window-azure-stream-analytics), you can apply a snapshot window by adding System.Timestamp() to the GROUP BY clause.

Exam Prep:

Azure Stream Analytics

1. Window functions: You must know the practical difference between all the stream analytics windowing functions & their usage (Tumbling, Hopping, Sliding & Session windows). hands-on
2. Learn how to use lookup data in the Azure Stream Analytics in a data streaming pipeline. hands-on
3. Azure Stream Analytics on IoT Edge

[GE] You develop data engineering solutions for a company. You need to ingest and visualize real-time Twitter data by using Microsoft Azure. Which three technologies should you use? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

A. Event Grid topic

**B. Azure Stream Analytics Job that queries Twitter data from an Event Hub**

C. Azure Stream Analytics Job that queries Twitter data from an Event Grid

**D. Logic App that sends Twitter posts which have target keywords to Azure**

E. Event Grid subscription

**F. Event Hub instance**

[GE] You are developing a solution that will stream to Azure Stream Analytics. The solution will have both streaming data and reference data. Which input type should you use for the reference data?

A. Azure Cosmos DB

B. Azure Event Hubs

**C. Azure Blob storage**

D. Azure IoT Hub

*See above diagram*

[GE] You use Azure Stream Analytics to receive Twitter data from Azure Event Hubs and to output the data to an Azure Blob storage account. You need to output the count of tweets during the last five minutes every five minutes. Each tweet must only be counted once. Which windowing function should you use?

A. a five-minute Session window

B. a five-minute Sliding window

**C. a five-minute Tumbling window**

D. a five-minute Hopping window that has one-minute hop

Tumbling window functions are used to segment a data stream into distinct time segments and perform a function against them, such as the example below. The key differentiators of a Tumbling window are that they repeat, do not overlap, and an event cannot belong to more than one tumbling window.

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions>

[GE] A company has a real-time data analysis solution that is hosted on Microsoft Azure. The solution uses Azure Event Hub to ingest data and an Azure Stream Analytics cloud job to analyze the data. The cloud job is configured to use 120 Streaming Units (SU). You need to optimize performance for the Azure Stream Analytics job. Which two actions should you perform?

A. Implement event ordering

**B. Scale the SU count for the job up**

C. Implement Azure Stream Analytics user-defined functions (UDF)

D. Scale the SU count for the job down

E. Implement query parallelization by partitioning the data output

**F. Implement query parallelization by partitioning the data input**

Scale out the query by allowing the system to process each input partition separately.

A Stream Analytics job definition includes inputs, a query, and output.

Inputs are where the job reads the data stream from.

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-parallelization>

*Repartitioning scales Azure Stream Analytics queries*

*SU’s affect physical resources needed by each partition*

[GE] You have an Azure Stream Analytics query. The query returns a result set that contains 10,000 distinct values for a column named clusterID. You monitor the Stream Analytics job and discover high latency. You need to reduce the latency. Which two actions should you perform?

A. Add a pass-through query.

B. Add a temporal analytic function.

**C. Scale out the query by using PARTITION BY**

D. Convert the query to a reference query.

**E. Increase the number of streaming units.**

[GE] You are developing a solution that will use Azure Stream Analytics. The solution will accept an Azure Blob storage file named Customers. The file will contain both in-store and online customer details. The online customers will provide a mailing address. You have a file in Blob storage named LocationIncomes that contains median incomes based on location. The file rarely changes. You need to use an address to look up a median income based on location. You must output the data to Azure SQL Database for immediate use and to Azure Data Lake Storage Gen2 for long-term retention. Solution: You implement a Stream Analytics job that has one streaming input, one query, and two outputs. Does this meet the goal?

**No. We need one reference data input for LocationIncomes, which rarely changes.**

**Note: Stream Analytics also supports input known as reference data. Reference data is either completely static or changes slowly.**

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-add-inputs#stream-and-reference-inputs>

[GE] You are developing a solution that will use Azure Stream Analytics. The solution will accept an Azure Blob storage file named Customers. The file will contain both in-store and online customer details. The online customers will provide a mailing address. You have a file in Blob storage named LocationIncomes that contains median incomes based on location. The file rarely changes. You need to use an address to look up a median income based on location. You must output the data to Azure SQL Database for immediate use and to Azure DataLake Storage Gen2 for long-term retention.

Solution: You implement a Stream Analytics job that has one streaming input, one reference input, two queries, and four outputs. Does this meet the goal?

**Yes. We need one reference data input for LocationIncomes, which rarely changes. We need two queries, on for in-store customers, and one for online customers.For each query two outputs is needed.Note: Stream Analytics also supports input known as reference data. Reference data is either completely static or changes slowly.**

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-add-inputs#stream-and-reference-inputshttps://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-define-outputs>

**INPUT:** Streaming

Customer file

**QUERY**

1. in-store customers

**OUTPUT**

1. Azure SQL DB

2. Azure Datalake

**INPUT: Reference**

LocationIncomes

**QUERY**

2. online customers

**OUTPUT**

3. Azure SQL DB

4. Azure Datalake

[GE] You have an Azure Stream Analytics job. You need to ensure that the job has enough streaming units provisioned. You configure monitoring of the SU% Utilization metric. Which two additional metrics should you monitor?

A. Watermark Delay

**B. Late Input Events**

C. Out of order Events

**D. Backlogged Input Events**

E. Function Events

Late Input Events: events that arrived later than the configured late arrival tolerance window

In job diagram, there is a per partition backlog event metric for each input. If the backlog event metric keeps increasing, it’s also an indicator that the system resource is constrained (either because of output sink throttling, or high CPU).

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-scale-jobsMonitor>

[GE] Your company uses Azure Stream Analytics to monitor devices. The company plans to double the number of devices that are monitored. You need to monitor a Stream Analytics job to ensure that there are enough processing resources to handle the additional load. Which metric should you monitor?

A. Input Deserialization Errors

B. Early Input Events

C. Late Input Events

**D. Watermark delay**

There are several other resource constraints that can cause the streaming pipeline to slow down. The watermark delay metric can rise due to:

- Not enough processing resources in Stream Analytics to handle the volume of input events.

- Not enough throughput within the input event brokers, so they are throttled.

- Output sinks are not provisioned with enough capacity, so they are throttled.

The possible solutions vary widely based on the flavor of output service being used. <https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-time-handling>

[GE] You need to implement event processing by using Stream Analytics to produce consistent JSON documents. Which three actions should you perform?

A. Define an output to Cosmos DB.

B. Define a query that contains a JavaScript user-defined aggregates (UDA) function.

C. Define a reference input.

**D. Define a transformation query.**

**E. Define an output to Azure Data Lake Storage Gen2.**

**F. Define a stream input.**

DOCDB stored documents that connect to the sales data in SALESDB. The documents are stored in two different JSON formats based on the sales channel. The sales data including the documents in JSON format, must be gathered as it arrives and analyzed online by using Azure Stream Analytics. The analytic process will perform aggregations that must be done continuously, without gaps, and without overlapping. As they arrive, all the sales documents in JSON format must be transformed into one consistent format. Manage and develop data processing Question Set 1

**INPUT:** Streaming

*Define a stream input*

**QUERY**

*Define a transformation query.*

**OUTPUT**

*Define an output to Azure Data Lake Storage Gen2*

Watermark: An event time marker that indicates up to what point events have been ingressed to the streaming processor. Watermarks let the system indicate clear progress on ingesting the events. By the nature of streams, the incoming event data never stops, so watermarks indicate the progress to a certain point in the stream.

The watermark concept is important. Watermarks allow Stream Analytics to determine when the system can produce complete, correct, and repeatable results that don’t need to be retracted. The processing can be done in a predictable and repeatable way. For example, if a recount needs to be done for some error handling condition, watermarks are safe starting and ending points.

[GE] You need to implement complex stateful business logic within an Azure Stream Analytics service. Which type of function should you create in the Stream Analytics topology?

A. JavaScript user-define functions (UDFs)

B. Azure Machine Learning

**C. JavaScript user-defined aggregates (UDA)**

Azure Stream Analytics supports user-defined aggregates (UDA) written in JavaScript, it enables you to implement complex stateful business logic. Within UDA you have full control of the state data structure, state accumulation, state decumulation, and aggregate result computation.

<https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-javascript-user-defined-aggregates>

[GE] You need to process and query ingested Tier 9 data. Which two options should you use?

A. Azure Notification Hub

B. Transact-SQL statements

C. Azure Cache for Redis

D. Apache Kafka statements

**E. Azure Event Grid**

**F. Azure Stream Analytics**

Event Hubs provides a Kafka endpoint that can be used by your existing Kafka based applications as an alternative to running your own Kafka cluster. You can stream data into Kafka-enabled Event Hubs and process it with Azure Stream Analytics, in the following steps:

Create a Kafka enabled Event Hubs namespace.

Create a Kafka client that sends messages to the event hub.

Create a Stream Analytics job that copies data from the event hub into an Azure blob storage.

Scenario:

Tier 9 reporting must be moved to Event Hubs, queried, and persisted in the same Azure region as the company’s main office

<https://docs.microsoft.com/en-us/azure/event-hubs/event-hubs-kafka-stream-analytics>

[GE] You develop data engineering solutions for a company.

A project requires analysis of real-time Twitter feeds. Posts that contain specific keywords must be stored and processed on Microsoft Azure and then displayed by using Microsoft Power BI. You need to implement the solution.

Which five actions should you perform in sequence?

**[1] Create an HDInisght cluster with the Spark cluster type**

**[2] Create a Jyputer Notebook**

**[3] Create a table - The Jupyter Notebook that you created in the previous step includes code to create an hvac table.**

**[4] Run a job that uses the Spark Streaming API to ingest data from Twitter**

**[5] Load the hvac table into Power BI Desktop**

*You use Power BI to create visualizations, reports, and dashboards from the Spark cluster data.*

[GE] You develop data engineering solutions for a company.

You need to deploy a Microsoft Azure Stream Analytics job for an IoT solution. The solution must:

- Minimize latency.

- Minimize bandwidth usage between the job and IoT device.

Which four actions should you perform in sequence?

**[1] Create and IoT hub and add the Azure Stream Analytics module to the IoT Hub namespace**

**[2] Create an Azure Blob Storage container**

**[3] Create an Azure Stream Analytics edge job and configure job definition save location**

**[4] Configure routes -  
You are now ready to deploy the Azure Stream Analytics job on your IoT Edge device.**

# Lesson 16

Data Processing

Lamda Architecture:

<https://sqlwithmanoj.com/2018/02/16/what-is-lambda-architecture-and-what-azure-offers-with-its-new-cosmos-db/>

Lambda architecture is a data-processing architecture designed to handle massive quantities of data by taking advantage of both batch processing and stream processing methods, and minimizing the latency involved in querying big data.

It is a Generic, Scalable, and Fault-tolerant data processing architecture to address batch and speed latency scenarios with big data and map-reduce. It consists of three layers:

Batch Layer: has a master dataset (immutable, append-only set of raw data) and pre-computes the batch views.

Speed Layer: has Batch views for fast queries.

Service Layer:

Diagram

Description automatically generated

# Lesson 17

Monitoring

|  |  |  |
| --- | --- | --- |
| Technology | Monitoring available | Description |
| SQL Database |  | The **sys.resource\_stats** returns historical data for CPU, IO, DTU consumption. |
| Azure SQL Database |  | Create a new action group  Use all security operations as a condition. Use all Azure SQL Database servers as a resource *Azure SQL Data Warehouse cache must be monitored* |
| Azure Synapse Analytics | Log files | On the master database, execute a query against the sys.dm\_pdw\_nodes\_os\_performance\_counters dynamic management view. |
| Data Warehouse (Aure Data Lake Gen1) | SqlRequests | **Diagnostic logging**  Configure Azure Data Lake Storage diagnostics to store logs and metrics in a storage account. |
| HDInsight cluster | Ambari REST API  Ambari Web UI | Monitor health of a HDInsight cluster (monitoring utilization across the whole cluster) |
|  | Azure Log Analytics | Give info on how to scale clusters Log data can be aggregated,  and a Log Analytics Workspace will store the collected data |
|  | HDInsight cluster management solutions | Cluster-specific management solutions that you can add for Azure Monitor logs |
|  |  |  |

[GE] You have the Diagnostics settings of an Azure Storage account as shown in the following exhibit. How long will the logging data be retained?

**A. 7 days**

B. 365 days

C. indefinitely

D. 90 days

[GE] You manage a process that performs analysis of daily web traffic logs on an HDInsight cluster. Each of the 250 web servers generates approximately 10 megabytes (MB) of log data each day. All log data is stored in a single folder in Microsoft Azure Data Lake Storage Gen 2. You need to improve the performance of the process

**A. Combine the daily log files for all servers into one file**

**C. Move the log files into folders so that each day’s logs are in their own folder**

For Hive workloads, partition pruning of time-series data can help some queries read only a subset of the data which improves performance

[GE] You need to implement diagnostic logging for Data Warehouse monitoring. Which log should you use?

A. RequestSteps

B. DmsWorkers

**C. SqlRequests**

D. ExecRequests

*The Azure SQL Data Warehouse cache must be monitored when the database is being used.*

[GE] Contoso, Ltd. plans to configure existing applications to use Azure SQL Database. When security-related operations occur, the security team must be informed. You need to configure Azure Monitor while minimizing administrative effort. Which three actions should you perform? Each correct answer presents part of the solution. NOTE: Each correct selection is worth one point.

**A. Create a new action group to email alerts@contoso.com.**

B. Use alerts@contoso.com as an alert email address.

**C. Use all security operations as a condition.**

**D. Use all Azure SQL Database servers as a resource.**

E. Query audit log entries as a condition.

[GE] You develop data engineering solutions for a company. A project requires the deployment of resources to Microsoft Azure for batch data processing on Azure HDInsight. Batch processing will run daily and must:

- Scale to minimize costs

- Be monitored for cluster performance. You need to recommend a tool that will monitor clusters and provide information to suggest how to scale.

Solution: Monitor clusters by using Azure Log Analytics and HDInsight cluster management solutions. Does the solution meet the goal?

**Yes. HDInsight provides cluster-specific management solutions that you can add for Azure Monitor logs. Management solutions add functionality to Azure Monitor logs, providing additional data and analysis tools. These solutions collect important performance metrics from your HDInsight clusters and provide the tools to search the metrics. These solutions also provide visualizations and dashboards for most cluster types supported in HDInsight. By using the metrics that you collect with the solution, you can create custom monitoring rules and alerts**

[GE] You develop data engineering solutions for a company. A project requires the deployment of resources to Microsoft Azure for batch data processing on Azure HDInsight. Batch processing will run daily and must:

- Scale to minimize costs

- Be monitored for cluster performance

You need to recommend a tool that will monitor clusters and provide information to suggest how to scale. Solution: Monitor cluster load using the Ambari Web UI. Does the solution meet the goal?

**No. Ambari Web UI does not provide information to suggest how to scale. Instead monitor clusters by using Azure Log Analytics and HDInsight cluster management solutions**

[GE] You manage a solution that uses Azure HDInsight clusters. You need to implement a solution to monitor cluster performance and status. Which technology should you use?

A. Azure HDInsight .NET SDK

B. Azure HDInsight REST API

C. Ambari REST API

D. Azure Log Analytics

**E. Ambari Web UI**

Ambari is the recommended tool for monitoring utilization across the whole cluster. The Ambari dashboard shows easily glanceable widgets that display metrics such as CPU, network, YARN memory, and HDFS disk usage. The specific metrics shown depend on cluster type. The “Hosts” tab shows metrics for individual nodes so you can ensure the load on your cluster is evenly distributed. The Apache Ambari project is aimed at making Hadoop management simpler by developing software for provisioning, managing, and monitoring Apache Hadoop clusters. Ambari provides an intuitive, easy-to-use Hadoop management web UI backed by its RESTful APIs.

[GE] A company has a Microsoft Azure HDInsight solution that uses different cluster types to process and analyze data. Operations are continuous. Reports indicate slowdowns during a specific time window. You need to determine a monitoring solution to track down the issue in the least amount of time. What should you use?

A. Azure Log Analytics log search query

**B. Ambari REST API**

C. Azure Monitor Metrics

D. HDInsight .NET SDK

E. Azure Log Analytics alert rule query

*Ambari is the recommended tool for monitoring the health for any given HDInsight cluster.*

[GE] You are monitoring an Azure Stream Analytics job. You discover that the Backlogged Input Events metric is increasing slowly and is consistently non-zero. You need to ensure that the job can handle all the events. What should you do?

A. Change the compatibility level of the Stream Analytics job.

**B. Increase the number of streaming units (SUs).**

C. Create an additional output stream for the existing input stream.

D. Remove any named consumer groups from the connection and use $default.

[GE] A company uses Azure Data Lake Gen 1 Storage to store big data related to consumer behavior. You need to implement logging. Solution: Use information stored in Azure Active Directory reports. Does the solution meet the goal?

**No. Instead configure Azure Data Lake Storage diagnostics to store logs and metrics in a storage account.**

[GE] A company uses Azure Data Lake Gen 1 Storage to store big data related to consumer behavior. You need to implement logging. Solution: Create an Azure Automation runbook to copy events. Does the solution meet the goal?

**No. Instead configure Azure Data Lake Storage diagnostics to store logs and metrics in a storage account.**

[GE] You need setup monitoring for tiers 6 through 8. What should you configure?

A. extended events for average storage percentage that emails data engineers

B. an alert rule to monitor CPU percentage in databases that emails data engineers

C. an alert rule to monitor CPU percentage in elastic pools that emails data engineers

D. an alert rule to monitor storage percentage in databases that emails data engineers

**E. an alert rule to monitor storage percentage in elastic pools that emails data engineers**

Scenario: Tiers 6 through 8 must have unexpected resource storage usage immediately reported to data engineers. Tier 3 and Tier 6 through Tier 8 applications must use database density on the same server and Elastic pools in a cost-effective manner

[GE] How should you monitor SALESDB to meet the technical requirements?

**A. Query the sys.resource\_stats dynamic management view.**

B. Review the Query Performance Insights for SALESDB.

C. Query the sys.dm\_os\_wait\_stats dynamic management view.

D. Review the auditing information of SALESDB.

Scenario: Disk IO, CPU, and memory usage must be monitored for SALESDB. The sys.resource\_stats returns historical data for CPU, IO, DTU consumption.

<https://dataplatformlabs.com/monitoring-azure-sql-database-with-sys-resource_stats/>

[GE] Your company uses several Azure HDInsight clusters. The data engineering team reports several errors with some applications using these clusters. You need to recommend a solution to review the health of the clusters. What should you include in your recommendation?

A. Azure Automation

**B. Log Analytics**

C. Application Insights

Azure Monitor logs integration. Azure Monitor logs enables data generated by multiple resources such as HDInsight clusters, to be collected and aggregated in one place to achieve a unified monitoring experience. As a prerequisite, you will need a Log Analytics Workspace to store the collected data. If you have not already created one, you can follow the instructions for creating a Log Analytics Workspace. You can then easily configure an HDInsight cluster to send many workload-specific metrics to Log Analytics.

<https://azure.microsoft.com/sv-se/blog/monitoring-on-azure-hdinsight-part-2-cluster-health-and-availability/>

[GE] You have an enterprise data warehouse in Azure Synapse Analytics named DW1 on a server named Server1. You need to verify whether the size of the transaction log file for each distribution of DW1 is smaller than 160 GB. What should you do?

**A. On the master database, execute a query against the sys.dm\_pdw\_nodes\_os\_performance\_counters dynamic management view.**

B. From Azure Monitor in the Azure portal, execute a query against the logs of DW1.

C. On DW1, execute a query against the sys.database\_files dynamic management view.

D. Execute a query against the logs of DW1 by using the Get-AzOperationalInsightSearchResult PowerShell cmdlet.

The following query returns the transaction log size on each distribution. If one of the log files is reaching 160 GB, you should consider scaling up your instance or limiting your transaction size.

-- Transaction log size

SELECT instance\_name as distribution\_db, cntr\_value\*1.0/1048576

as log\_file\_size\_used\_GB, pdw\_node\_id

FROM sys.dm\_pdw\_nodes\_os\_performance\_counters

WHERE instance\_name like 'Distribution\_%'

AND counter\_name = 'Log File(s) Used Size (KB)'

<https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-manage-monitor>

[GE] You implement an event processing solution using Microsoft Azure Stream Analytics.

The solution must meet the following requirements:

- Ingest data from Blob storage

- Analyze data in real time

- Store processed data in Azure Cosmos DB

Which three actions should you perform in sequence?

**[1] Configure Blob storage as input; select items with the TIMESTAMP BY clause**

**[2] Set up cosmos DB as the output**

**[3] Create a query statement with the SELECT INTO statement.**

Exam Prep:

1. Understand the SQL auditing features & do a hand-on experiment on who/when & what got accessed from the Azure SQL DB & WH? hands-on
2. Learn about enabling SQL server automatic tuning & give special attention towards the inheritance with tuning options like Force Plan, Create Index & Drop Index. hands-on
3. Read & understand In-Memory technologies that improve performance without increasing the database service tier. hands-on
4. Understand the materialized view design pattern and think about its uses to boost a slow-performing SQL query. hands-on
5. Learn how to enable and configure logging of diagnostics telemetry for Azure SQL databases. Pay special attention to the metric storage options like Azure SQL Analytics, Azure Event Hubs & Azure Storage. hands-on
6. Imbibe ADLS Gen2 performance optimization techniques. Understand file sizing & data organization into folders.
7. Discover Azure data factory monitoring using Azure monitor and think about use cases like last quarter log analytics and find out different actionable trends. hands-on
8. Learn about on-premises HA data gateway cluster to avoid single points of failure and to load balance traffic across gateways in a cluster.
9. Understand the use of Azure Log Analytics to monitor HDInsight clusters. Pay special attention to “Install HDInsight cluster management solutions”. hands-on

# Lesson 18

CosmosDB

Azure Cosmos containers

<https://docs.microsoft.com/en-us/azure/cosmos-db/databases-containers-items>

An Azure Cosmos container is the unit of scalability both for provisioned throughput and storage. A container is horizontally partitioned and then replicated across multiple regions. The items that you add to the container are automatically grouped into logical partitions, which are distributed across physical partitions, based on the partition key. The throughput on a container is evenly distributed across the physical partitions

API-specific entities

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Entity | CosmosDB | SQL | MongoDB | Table | Cassandra | Gremlin |
| Database |  | Database | Database | N/A | Keyspace | Database |
| Container | Azure Cosmos container | Container | Collection | Table | Table | Graph |
| Item | Azure cosmos item | Document | Document | Item | Row | Node or Edge |

[GE] You are the data engineer for your company. An application uses a NoSQL database to store data. The database uses the key-value and wide-column NoSQL database type. Developers need to access data in the database using an API. You need to determine which API to use for the database model and type. Which two APIs should you use?

A. Table API

**B. MongoDB API**

C. Gremlin API

D. SQL API

**E. Cassandra API**

Azure Cosmos DB is the globally distributed, multimodel database service from Microsoft for mission-critical applications. It is a multimodel database and supports document, key-value, graph, and columnar data models.

Wide-column stores store data together as columns instead of rows and are optimized for queries over large datasets. The most popular are Cassandra andHBase. <https://docs.microsoft.com/en-us/azure/cosmos-db/graph-introductionhttps://www.mongodb.com/scale/types-of-nosql-databases>

[GE] You are developing a data engineering solution for a company. The solution will store a large set of key-value pair data by using Microsoft Azure Cosmos DB. The solution has the following requirements:

- Data must be partitioned into multiple containers.

- Data containers must be configured separately.

- Data must be accessible from applications hosted around the world.

The solution must minimize latency. You need to provision Azure Cosmos DB.

Which three actions should you perform?

A. Configure account-level throughput.

B. Provision an Azure Cosmos DB account with the Azure Table API. Enable geo-redundancy.

C. Configure table-level throughput.

D. Replicate the data globally by manually adding regions to the Azure Cosmos DB account.

**E. Provision an Azure Cosmos DB account with the Azure Table API. Enable multi-region writes.**

Table API for applications that are written for Azure Table storage and that need premium capabilities like:

Turnkey global distribution.

Dedicated throughput worldwide (when using provisioned throughput).

Single-digit millisecond latencies at the 99th percentile.

Guaranteed high availability.

Automatic secondary indexing.

Scale read and write throughput globally. You can enable every region to be writable and elastically scale reads and writes all around the world. The throughput that your application configures on an Azure Cosmos database or a container is guaranteed to be delivered across all regions associated with your Azure Cosmos account. The provisioned throughput is guaranteed up by financially backed SLAs.

<https://docs.microsoft.com/en-us/azure/cosmos-db/distribute-data-globally>

**Case Study 1**

Most of the company’s data reside in Microsoft SQL Server database. Application databases fall into one of the following tiers:

|  |  |  |  |
| --- | --- | --- | --- |
| Applications | Tier | Replication | Notes |
| Internal Contoso | 1 | Yes |  |
| Internal Contoso | 2 | SQL Data Sync | Data Sync between databases |
| Internal Partner | 3 | Yes | Replicate to Partner |
| External Contoso | 4,5,6 | Yes |  |
| External Partner | 7,8 | No | Partner managed |
| Internal Distribution and Sales | 9 | Yes, once ingested at branches | Data ingested from Contoso branches |
| External Distribution and Sales | 10 | Yes, once ingested at Contoso main office | Data is ingested from multiple sources |

Monitoring must be set up on every database. Contoso and partners must receive performance reports as part of contractual agreements. Tiers 6 through 8 must have unexpected resource storage usage immediately reported to data engineers.

The Azure SQL Data Warehouse cache must be monitored when the database is being used. A dashboard monitoring key performance indicators (KPIs) indicated

by traffic lights must be created and displayed based on the following metrics:

|  |  |
| --- | --- |
| Metric | Description |
| A | Low cache hit %, high cache usage % |
| B | Low cache hit %, low cache usage % |
| C | High cache hit %, high cache usage % |

Requirements:

You identify the following reporting requirements:

Azure Data Warehouse must be used to gather and query data from multiple internal and external databases

Azure Data Warehouse must be optimized to use data from a cache

Reporting data aggregated for external partners must be stored in Azure Storage and be made available during regular business hours in the connecting regions

Reporting strategies must be improved to real time or near real time reporting cadence to improve competitiveness and the general supply chain

Tier 9 reporting must be moved to Event Hubs, queried, and persisted in the same Azure region as the company’s main office

Tier 10 reporting data must be stored in Azure Blobs

Issues:

Both internal and external client application run complex joins, equality searches and group-by clauses. Because some systems are managed externally, the

queries will not be changed or optimized by Contoso

External partner organization data formats, types and schemas are controlled by the partner companies

Internal and external database development staff resources are primarily SQL developers familiar with the Transact-SQL language.

Size and amount of data has led to applications and reporting solutions not performing are required speeds

Tier 7 and 8 data access is constrained to single endpoints managed by partners for access

The company maintains several legacy client applications. Data for these applications remains isolated form other applications. This has led to hundreds of

databases being provisioned on a per application basis

**Case Study 2**